Investigating the feasibility of crisis-discharge decision-support to reduce readmission rates at a psychiatric ward.

Marelise Hattingh

Department of Industrial Engineering
Stellenbosch University

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering in the Faculty of Engineering at Stellenbosch University.

M.Eng. (Research) Industrial

Supervisor: Louzanne Bam
Co-supervisor: Prof. Jan H. van Vuuren

December 2016

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work; that I am the sole author thereof (save to the extent explicitly otherwise stated); that reproduction and publication thereof by Stellenbosch University will not infringe any third-party rights; and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2016

Abstract

The pressure on the availability of beds in South African psychiatric hospitals is high. In response, the Western Cape has implemented a "crisis" discharge policy. This policy is different from the planned short stay practice as the patients are discharged earlier than what is clinically ideal to allow people on the waiting list to be admitted. The crisis-discharge policy therefore strives to optimise the combined healthcare outcome for the patient population (including those currently in the ward and those on the waiting list for admission). Discharge and admission decisions are informed by reviewing clinical indicators of the patient population.

A previous study conducted at Stikland Psychiatric Hospital, which is also the institution where this research was undertaken, reported crisis-discharge to be a significant predictor of an increased risk for readmission. This suggests that implementing a crisis-discharge policy to alleviate the pressure on available beds, may in fact exacerbate the scenario. Currently, unaided decision making is implemented by the clinical psychiatrists to solve this combinatorial optimisation problem and it is therefore unlikely that the daily decisions are optimal.

This study investigates readmission at Stikland Psychiatric Hospital, specifically at the acute male inpatient ward to (i) determine whether variables exist that indicate that certain patients within this population have a higher risk of requiring readmission after a crisis-discharge; and, if such variables do exist, (ii) to determine the predictive capability of these variables with a view of recommending the feasibility of a decision-support system for crisis-discharge at the male inpatient ward. Various patient variables such as age, diagnosis, place of follow-up and substance use are analysed.

Basic descriptive methods, biostatistics and data mining were employed to analyse the data. Predictive models were developed and evaluated using, amongst others, classification and regression trees and random forests. The research was conducted with continuous input from clinical subject matter experts.

The most important statistically significant variables pertaining to the risk of readmission are the diagnosis, whether a patient belongs to a community after-care programme, and the area that a patient originates from. Direct admissions and schizophrenic patients were found to be twice as likely to be readmitted as patients who are not from these groups. The schizo-affective and bipolar diagnostic groups are about three times as likely to be readmitted compared to patients who are not from these diagnostic groups. The substance induced psychosis diagnostic variable, and a community programme variable indicated that patients were less than half as likely to require readmission. These results are some of the insights that are presented in this research project.

The best-performing predictive model is able to classify whether patients would require readmission following a crisis-discharge with average accuracy of 70%. Based on these findings, further research towards to the development of a crisis-discharge decision-support tool is recommended.

Opsomming

Die aanvraag na beskikbare beddens in Suid-Afrikaanse psigiatriese hospitale is hoog. Dit het gelei tot die implementering van 'n krisis-ontslag beleid in die Wes-Kaap. Hierdie beleid verskil van die beplande korter-hospitaalvertoeftyd aangesien 'n pasiënt vroeër ontslaan word as wat klinies ideaal is sodat 'n persoon op die waglys toegelaat kan word. Die krisis-ontslag beleid poog dus om die gekombineerde gesondheidsorg uitkoms van die totale pasiëntpopulasie te optimeer (dit sluit die pasiënte in die saal asook die op die waglys in). Besluite rakende die ontslaan en toelating van pasiënte word geneem deur kliniese veranderlikes van die pasiëntpopulasie in ag te neem.

'n Vorige studie gedoen by die Stikland Psigiatriese Hospitaal, dieselfde instansie waarby hierdie navorsing gedoen is, het bevind dat krisis-ontslag 'n beduidende bepaler van hertoelating is. Dit dui daarop dat 'n krisis-ontslag beleid nie die druk op die aantal beskikbare beddens verlig nie, maar dit egter kan vererger. Tans word die kombineerde optimaliseringsprobleem opgelos deur die kliniese psigiaters wat op 'n daaglikse basis besluite moet neem, en gevolglik is dit onwaarskynlik dat die besluite wel optimaal is.

Hierdie navorsingsprojek ondersoek hertoelating, spesifiek by die akute manssaal van Stikland Psigiatriese Hospitaal, om vas te stel of (i) daar veranderlikes bestaan wat daarop dui dat sekere pasiënte van hierdie populasie 'n hoër risiko het om hertoegelaat te word na 'n krisis-ontslag, en, indien diesulke veranderlikes voorkom, (ii) wat die voorspellingsvermoë van hierdie veranderlikes is om sodoende die moontlikheid van 'n besluitsteunstelsel vir krisis-ontslag vas te stel. Pasiënt veranderlikes wat ondersoek word is onder andere ouderdom, diagnose, opvolg en dwelmgebruik.

Die data is geanaliseer met beskrywende statistiek, verskeie biostatistiese en *data mining* metodes. Klassifikasie en regressie bome en *random forests* is onder andere gebruik om voorspellingsmodelle te ontwikkel en te evalueer. Die navorsing het deurlopende kontaksessies met kliniese kundiges behels.

Die belangrikste statisties beduidende veranderlikes, wat die risiko vir hertoelating benadruk, is die gebied vanwaar 'n pasiënt kom, diagnose en of 'n pasiënt aan 'n gemeenskapsopvolg-program behoort. Direkte toelatings– en skisofreniese pasiënte het 'n twee maal groter waarskynlikheid vir hertoelating as pasiënte wat nie onder hierdie twee groepe klassifiseer nie. Die skiso-affektiewe– en bipolêre diagnose groep is ongeveer drie keer meer waarskynlik vir hertoelating as pasiënte wat nie in hierdie diagnose-groepe val nie. Daar is ook gevind dat die substansgeïnduseerde psigose veranderlike en pasiënte wat behoort aan 'n spesifieke gemeenskapsopvolg-program meer as die helfte minder kans het om hertoegelaat te word. Hierdie resultate verteenwoordig van die bevindinge wat bespreek word in hierdie navorsingsprojek.

Die beste voorspellingsmodel het met 'n gemiddelde akkuraatheid van 70% voorspel of krisis-ontslag pasiënte weer hertoegelaat gaan word. Hierdie bevindinge lei dan tot die aanbeveling dat verdere navorsing gedoen moet word vir die ontwikkeling van 'n krisis-ontslag besluitsteun-program.

Acknowledgements

The author wishes to acknowledge the following people for their various contributions towards the completion of this project:

My supervisor, Louzanne Bam, for her guidance, perspective and encouragement during this project.

The CSIR, for the opportunity and financial support to further my studies in the area that interest me.

Dr. Ingé-Marli Smit and Prof. Liezl Koen, without whom this project would not have been possible, for providing advice, offering help with understanding the data as well as motivating along the way. I respect and find inspiration in the medical doctors who make time to conduct research because,

"When you treat a patient, you treat one patient. When you do research, you treat 10 000 patients." –R.H. Riffenburgh

My parents who have helped me throughout my life and this project, by providing valuable input, unconditional support and regular sanity checks.

My friends, who are by now regarded family and have always believed in me more than I believe in myself.

Most importantly, God, who provides courage, energy, peace and a clear mind.

Table of Contents

Nomenclature

ACRONYMS

ACT	Assertive Community Treatment
ADHD	Attention-deficit hyperactivity disorder
AUC	Area under curve
CART	Classification and regression trees
CBS	Community based service
CI	Confidence interval
CRISP-DM	Cross industry standard process for data mining
CV	Cross validation
dof	Degrees of freedom
DoH	Department of Health
EMS	Emergency medical service
GMC	General medical condition
GDP	Gross domestic product
HIV	Human immunodeficiency virus
IDS	Intellectual disability patient
KDD	Knowledge discovery in databases
kNN	k-nearest neighbours
LOS	Length of stay
MDD	Major depressive disorder
MDGs	Millennium Development Goals
MHAP	Mental health action plan

MHCA	Mental Health Care Act No. 17
MHPF	Mental Health Policy Framework
MLP	Multilayer perceptron
MLPN	Multilayer perceptron network
M/SUD	Mood and substance abuse disorders
NCS	National core standards
NDP	National development plan
NN	Neural networks
OECD	Organisation for Economic Co-operation and Development
PHC	Primary healthcare
S	Schizophrenia
SA	Schizo-affective
SAFMH	South African Federation for Mental Health
SEMMA	Sample, explore, modify, model and assess
SIPD	Substance induced psychotic disorder
SME	Subject matter expert
SVM	Support vector machines
TB	Tuberculosis

GREEK SYMBOLS

α	The level of significance for a test
λ	The ratio of variability in a regression model
μ	Population mean
$\mu\|x$	Population mean given a x value
ρ	Population correlation coefficient
τ	Node of a decision tree
τ_L	Left daughter node of a decision tree
τ_R	Right daughter node of a decision tree
ϕ	Symmetric function for all splitting probabilities in a decision tree

Π_i	Class i of a CART tree

ROMAN SYMBOLS

A_i	Binary variable representing product i that is bought from a store
A_{milk}	Binary variable indicating whether milk is bought from a store or not
b	Regression coefficient for a regression model constructed from a sample
B	Population regression coefficient for a regression model
$\mathcal{B}_0$	Intercept of the slope-intercept regression model
$\mathcal{B}_1$	Slope of the simple regression model
C	Set of classes for a classification tree
$E(y\|x)$	Expected population value of y given x
F	The size of a random forest model
F^T	A random forest tree
$\mathbf{g}$	Measurement vector for a CART model
G_i	Split i for subset G (decision tree)
h	Variable in a random forest model
H	The set of variables included in a random forest model
H_0	The null hypothesis
H_1	The alternative hypothesis
$H(t)$	Hazard at time t
$H_0(t)$	Baseline hazard for readmission at time t
$i(\tau)$	Impurity of node τ
j	The class label of a terminal node
J	Amount of classes in variable
k	Amount of readmissions a subject has experienced
$\mathcal{L}$	The learning set
M	Categorical variable comprising of ℓ amount of categories
n	Total observations in a terminal node
n_s	Sample size

n_{min}	Minimum number of observations that should be in a terminal node
N	The total amount of observation in the learning set
p_m	Sample proportion
q	The total amount of classes in the learning set for a classification tree
r	Correlation coefficient
R^2	Coefficient of determination
$R(T)$	True misclassification rate of the tree classifier
$R^{Re}(T)$	Resubstitution estimate of the tree classifier
s_e	The standard deviation between the observed values and the regression line (residuals)
s_m	Standard error
m	Predictor variables in a Cox regression model
s_x	Standard deviation in the independent observations (x)
s_{xy}	The covariance of x and y
s_y	Standard deviation in the dependent observations (y)
$s_{y\|x}$	Standard deviation of the estimate of the individual predictions of y for a given x value
$s_{\bar{y}\|x}$	Estimation of the mean values of y given an independent (x) observation
t	Variable indicating the survival time of a subject
T	Tree classifier (prediction model)
$\tilde{T}$	Set of all terminal nodes T
$\mathcal{T}$	The test set
t_{crit}	Critical t-value from the t distribution
t_{start}	Starting time
t_{stop}	Starting time
t_{TS}	Test statistic from the t-table
v	The number of random, approximate equally sized sub-samples that are formed from the learning set applied in v-fold cross validation
$\mathbf{v}$	Measurement vector for CART example
V	Subset for all $\mathbf{v}$ measurement vectors

$\mathcal{V}$	The validation set
v_1 and v_2	Variables in CART example for blood pressure and age respectively
x_i	Variable representing independent observations in a data set
$\bar{x}$	The average of the independent data points
X	Independent variable vector
y_i	Variable representing the dependent outcome value for a dataset
$\bar{y}$	The average of the dependent data points
$\mathbf{y}$	Variable that is to be predicted
Y	Dependent variable vector
$y\|x$	Predictive strength of y given x

TERMINOLOGY

ACT/NB	A merged variable representing an observation where the patient simultaneously belongs to the New Beginnings and Assertive Community Treatment programmes.
ANOVA	The process of analysing variance.
Agoraphobia	The fear of open or public spaces.
Backpropagation	A common method of training artificial neural networks used in combination with an optimisation method.
Bed pressures	The pressure on hospitals to have space (beds) available for patients that require admission.
Bilateral donors	Government donors who provide assistance directly to a recipient country.
Bivariate analyses	Analysis of two variables, usually to determine the relationship between a dependent and independent variable.
Cannabis	A plant that is used as a drug and also referred to as 'dagga' or 'marijuana'.
Cognitive	Cognitive functions enable a human to reason, remember, pay attention and learn or talk a language, all leading to gaining information, and accordingly, knowledge.
Decentralisation	Transferring some psychiatric services that usually take place at a psychiatric hospital, to community-based programmes or institutions.

Deinstitutionalisation	The process of discharging patients to accelerate the transition from the psychiatric institution to the community.
Deming regression	Fits the best line for a two-dimensional dataset, but other than simple regression it accounts for error in both the dependent and independent observations.
Disability-adjusted life year	Is expressed as the number of years lost because of disability, early death or ill-health. It is a measure of the total disease burden.
Dysthymia	A medical term to describe re-occurring mild depression.
Generalist	A person with occupational health training.
Global Assessment of Functioning	A numeric scale that subjectively rate the occupational, psychological and social functioning of adults.
GMC_Other	Variable representing patients diagnosed with either a general medical condition or 'other' ('other' not being schizophrenia, schizo-affective, bipolar, substance induced psychotic disorder, major depressive disorder, anxiety, or a general medical condition).
Hazard ratio	An instantaneous event rate indicating the probability that an individual at time t experiences an event, assuming the patient is event-free (survived) up to time t.
Lifetime prevalence	The proportion of a population that will experience some condition at some point in their life.
MDD&Anxiety	Merged variable representing patients diagnosed with either major depressive disorder or anxiety.
Methamphetamine	A strong central nervous system stimulant used as a drug (also referred to as 'tik' in this research).
Multilayer perceptron network	A multilayer perceptron is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An multilayer perceptron network consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.
Non-communicable disease	Any illness or disease that cannot be transmitted to another person, e.g. heart disease and diabetes.
Nonparametric test	A hypothesis test conducted without the assumption that the distribution of the population is characterised by certain parameters.
Odds ratio	The odds that an outcome (readmission) will occur given a particular condition, compared to the odds of the outcome occurring in the absence of that condition.

Opium	An addictive narcotic drug.
Organic mental disorder	Decreased mental function because of a physical disease.
Psychiatrist	A medical doctor or physician who specialised further in the field of clinical psychotherapy.
Psychogeriatric services	Services that focus on patients with age-related mental impairment.
Psychosis	A severe mental disease in which contact is lost with reality owing to impaired thoughts and emotions.
Psychotherapy	The general term for receiving treatment from a mental health care provider for any mental health problems.
Psychotropic drugs	Treatment that can influence the brain function and is used to treat depression, insomnia and anxiety.
Residual	The difference between the observed value and the predicted value of the dependent variable.
SA_Bi_MDD&Anx	A variable group representing the diagnosis of a patient which is either schizo-affective, bipolar, major depressive disorder or anxiety.
Sigma Six	A data-driven approach for eliminating defects from manufacturing to product and service.
STL_NB_ACT	A merged variable group representing follow-up of a patient at either Stikland Psychiatric Hospital, New Beginnings or the Assertive Community Treatment programme.
Tik	A type of methamphetamine (substance) generally produced in makeshift laboratories.
Tyg_other_none	A variable group representing follow-up of a patient which is either at Tygerberg, no follow-up or 'other'.
Wald statistic	The outcome of a test which determines whether a variable contributes significantly to predicting the dependent variable.

List of Figures

List of Tables

CHAPTER 1

Introduction

The purpose of this document is to investigate whether readmission rates at an inpatient ward of a psychiatric hospital where a "crisis"-discharge policy is implemented, can be reduced by providing decision support on the likelihood of each patient requiring readmission if they were to be crisis-discharged.

This chapter serves as an introduction: background is given on the project and origin; the problem statement is defined along with the hypothesis, objectives, and ethical implications; the research design and methodology are presented; and the document's structure is described.

1.1 Project background and origin

The pressure on the amount of available beds in South African psychiatric hospitals is high. In the Western Cape, an early "crisis"-discharge policy has been implemented in psychiatric hospitals in order to deal with acute 'bed pressures' (Niehaus *et al.*, 2008). The early crisis-discharge policy implies that inpatients are discharged after a shorter length of stay than what is ideal to allow for individuals on the waiting list to be admitted to the ward. These decisions are made on a daily basis by a team of healthcare professionals after reviewing the clinical indicators of the current inpatients as well as individuals on the waiting list.

Internationally, there has been a decrease in the length of hospital stay of psychiatric inpatients over the past few decades, but this trend is explained by discharge planning with the goal of optimising the health benefit of each patient involved – regardless of capacity constraints. Crisis-discharge, in contrast, seeks to optimise the combined healthcare outcome for the current patients as well as those on the waiting list. A study conducted at the Stikland Psychiatric Hospital's acute male ward, in the Western Cape, found that the most significant indicator for an increased risk of readmission was indeed crisis-discharge (Niehaus *et al.*, 2008).

The existence of a crisis-discharge policy in the Western Cape indicates that there is insufficient capacity to serve the need for psychiatric inpatient care in the province. Based on the research by Niehaus *et al.* (2008) a crisis-discharge policy could cause bed pressure to increase rather than decrease, owing to crisis-discharged patients being more likely to require readmission.

1.2 Rationale of the study

In this section the problem statement, aim, objectives, and ethical implications are described.

1.2.1 Problem statement

Health care professionals have to make crisis-discharge as well as admission decisions on a daily basis. These types of decisions constitute a combinatorial optimisation problem. Currently, unaided decision-making is implemented at Stikland Psychiatric Hospital, and accordingly, there is no guarantee that the decisions made are, in fact, optimal. This research proposes using historical data from the male inpatient ward at Stikland Psychiatric Hospital to determine the probability of a patient requiring readmission should they be crisis-discharged. This information could be used to support psychiatrists' crisis-discharge decision-making with the goal of reducing the readmission rates.

1.2.2 Aim

The aim of this research is (i) to determine whether specific variables exist that indicate that certain patients within the Stikland male inpatient population would have a higher risk of requiring readmission after a crisis-discharge; and, if such variables do exist, (ii) to determine the likely predictive capability of these variables with a view of recommending the feasibility of a decision-support system for crisis-discharge at the Stikland male inpatient ward.

1.2.3 Objectives

The following research objectives have been identified in order to achieve the research aim:

- Document the current business processes and decision-making at Stikland Psychiatric Hospital;
- Determine the attributes of patients most likely chosen for crisis-discharge;
- Define and prioritise, based on analysis of the data, the factors that influence the risk of readmission;
- Develop and evaluate predictive models that can be used to calculate a patient's probability of readmission based on certain variables; and
- Discuss the feasibility of developing a decision support framework to assist healthcare professionals with the daily crisis-discharge decisions.

1.2.4 Ethical implications of the research

The problem and data are related to the healthcare sector and accordingly there are ethical aspects to consider. Sensitive patient data are handled and must be appropriately anonymised.

The model and methods should not be influenced by social factors and the capital worth of a person. Ethical approval was granted from the Health Research Ethical Committee of Stellenbosch University after which the National Health Research Database granted departmental and hospital approval.

Obtaining and cleaning the data was one of the most time consuming aspects of this research. Medical data is sensitive and case-specific, and, accordingly understanding the population variables and decisions is important. Throughout the research, especially during the data preparation phase and the interpretation of results, regular consultations took place with the psychiatrist at Stikland Psychiatric Hospital. The data preparation phase involved more than 20 contact sessions and regular email correspondence with the psychiatrists. It is imperative that the content and implications of the variables in the data set are understood by the researcher. The psychiatrist are seen as part of the research team and the analysis methodology and results were discussed and presented to the whole team. Any decisions with regard to grouping the data as well as interpreting the results were validated by the clinical research partners.

1.3 Proposed research design and project plan

In this section, the research design and research plan are described.

1.3.1 Research design

The study comprises various research techniques, but it is primarily an empirical study. Initially, qualitative research is conducted to document the current business and decision-making processes implemented at Stikland Psychiatric Hospital by conducting interviews with the psychiatrist. Non-empirical research is also done in the form of asking meta-analytic and theoretical questions. Information is gathered from existing literature pertaining to statistical methods and data mining as well as psychiatric care in South Africa.

The study becomes empirical in nature which involves observing (collecting and organising) basic empirical facts and through the process of induction, forming a hypothesis leading to deducting possible consequences from the empirical data, testing the hypothesis and evaluating the outcomes. Descriptive and inferential statistics along with data mining and survival analysis are used to analyse and describe the data, determine trends, and to build and evaluate predictive models.

1.3.2 Research plan

The methodology of this research predominantly constituted the following phases:

1. Initial meeting with the research team which included the psychiatrists;
2. A preliminary literature review to understand the real-world problem;
3. The project proposal – establish the hypothesis, scope, objectives, ethical considerations and research design;

4. Submit a protocol to gain ethical clearance for the project;
5. Conduct a literature review on:
   (a) South Africa's healthcare sector;
   (b) Psychiatric care in South Africa and the Western Cape;
   (c) Readmission, deinstitutionalisation and other phenomena in the psychiatric sector;
   (d) Investigating similar published studies that also investigated readmission; and
   (e) Introduction to data mining and common data mining methods.
6. Discuss the data analysis methodology with a statistical subject matter expert;
7. Interview the clinical subject matter expert to document the current discharge and admission procedures;
8. Consult with the clinical subject matter experts to discuss and understand the variables in the dataset;
9. Prepare and clean the dataset;
10. Decide on a viable software programme to analyse the data;
11. Continue and refine the literature review to include:
    (a) Introduction to statistical methods, especially regression models;
    (b) Predictive methods that can be used for this research; and
    (c) How to apply the statistical and data mining methods in the Statistica Software.
12. Perform basic statistical analysis to describe the variables in the dataset;
13. Conduct statistical and data mining analyses to determine the variables that have a significant influence on the probability of a patient being readmitted;
14. Evaluate and compare the results as well as the predictive capability of the models;
15. Build and evaluate predictive models which may be used to develop a decision support tool;
16. Present the final models to the statistical subject matter expert; and finally
17. Present the results to the clinical research partners.

Multiple consultations with both the clinical and statistical subject matter experts took place throughout this research project owing to the nature of the data being sensitive and case-specific.

1.4 Structure of the document

Chapter 2 provides a comprehensive overview of the real-world problem by (i) broadly describing the South African healthcare sector; (ii) providing background on psychiatric healthcare in South Africa and the Western Cape; (iii) describing the common phenomena of deinstitutionalisation and readmission in psychiatric hospitals; and (iv) presenting the problem in the context of the background provided in the chapter.

Chapter 3 gives an introduction to the science of 'learning from data'. The chapter entails a literature review on both statistical and data mining methods with focus given to methods that may be applicable to this research. Similar published studies that investigated readmission at various psychiatric institutions are introduced along with the methods used in their analyses. In Chapter 4, the variables of the similar published studies are presented whereafter the focus is shifted to this research and the process of developing a dataset suitable for analysis. The data learning methodology of this research is presented along with practically discussing the various descriptive and predictive methods that will be used to analyse the dataset.

Chapter 5 presents the results of the various descriptive and predictive analyses that were conducted. The variables and their relationship with readmission are discussed and compared with each analysis. The implication of the results are also discussed along with evaluating and comparing the various models' predictive capability. Chapter 6 summarises the research findings and discusses the results and implications from the real-world problem's perspective. The feasibility of further developing a decision support tool is discussed and a concluding overview of the research is presented.

1.5 Conclusion: Introduction

This chapter introduced the research project by briefly describing the problem, providing the rationale of the study along with the methodology and outlining the structure of the study.

In the following chapter, the real-world problem is described in more detail by introducing the South African healthcare sector and in particular psychiatric healthcare in the Western Cape. Psychiatric phenomena such as deinstitutionalisation and readmission are also described along with introducing published studies that are similar to this research project.

CHAPTER 2

The real-world problem

The project was introduced in Chapter 1 by presenting the research problem as well as defining the aim and objectives of the research. The research design and methodology followed by the structure of the document were also described.

In turn, this chapter comprises a literature review providing background information to which this project is related. It includes a section dedicated to the South African healthcare sector, describing some important legislation, statistics, the status quo and future goals. The psychiatric sector is investigated in a similar way by focusing on the Western Cape and then specifically Stikland Psychiatric Hospital. Mental illness is also defined along with providing background with regard to general phenomena in Mental Health Hospitals.

2.1 South African healthcare sector

There is a wide array of literature available on South Africa's healthcare sector, its various departments, budgets and spending, legislation, development plans and acts relating to the private and public health sector. This section provides a brief introduction on these topics to contextualise the problem under investigation.

2.1.1 Status quo

The South African healthcare system comprises a widely divided public and private sector in terms of quality and access (Ruff *et al.*, 2011). The public sector provides primary healthcare to citizens who do not have or cannot afford medical aid as well as to employees from the government, resulting in about 80% of the population (Human, 2010). According to the Health Minister, Aaron Motsolaedi (2013), 84% of the citizens have to do with "second-rate care" owing to not being able to afford private healthcare (News24, 2013). The public sector mainly comprises clinics in rural areas where the condition of the buildings and equipment are poor. People have trouble reaching the clinics in more isolated areas. The few academic hospitals situated in larger towns and cities offer better service, but struggle with capacity issues (Human, 2010).

The private sector, in turn, is known for providing greater quality care with improved accessibility. Medical staff are also more likely to move to the private sector owing to substantial better facilities and salaries (Human, 2010). According to Human (2010), one of South Africa's greatest challenges is to reduce the great difference in the quality of care provided by the private and public sectors. The private health sector also supports the economy by generating jobs and income, facilitating training, international programmes and creating investment opportunities (Econex, 2013).

In a business briefing in 2013, Motsoaledi claimed that a privileged few, 18% of the population, had access to private healthcare (News24, 2013). In fact, about 16% to 17% (2012–2013) of South Africa's population access private healthcare as members of medical aids (Econex, 2013). However, this can be misleading when one considers the people who make use of private healthcare, which includes general practitioners and dentists, and pay from their own funds, whereafter the percentage is estimated to be between 28% and 38% (Econex, 2013). In 2013, the private sector's expenditure made for about half of the national health expenditure, indicating the important role the sector plays in supporting the government in attaining the various health goals and providing in the basic rights of the population (Econex, 2013). The National Treasury estimated that the out-of-pocket spending on private healthcare amounted to about 17% and 15% of the total spending to the private sector in 2008 and 2012, respectively (Econex, 2013). In February 2016, an article was published on News24 that stated that according to the WHO, South Africa's private hospital costs are similar to that of countries such as the United Kingdom, France and Germany, which have much higher Gross Domestic Product (GDP) (Ngoepe, 2016). It also stated that 41.8% of the country's total healthcare expenditure is spent on private, voluntary health insurance, more than any of the Organisation for Economic Co-operation and Development[1] (OECD) countries (Ngoepe, 2016).

A study conducted by Econex in 2013 revealed a number of interesting findings regarding South Africa's healthcare environment. In 1989, 27% of the people belonging to a medical aid made use of public healthcare, in contrast, declining to 0.324% in 2013 (Econex, 2013). This may also indicate that the decline in the standard of care provided by provincial hospitals results in a higher demand for quality care from private institutions. In 2008 South Africa's private sector was ranked along with the likes of Australia, Belgium, Switzerland, Ireland and Sweden (Econex, 2013).

2.1.2 Public and national health programmes and legislation

Sections 27 and 28 of the South African constitution state that all South Africans have the right to healthcare access. The National Act, with the most recent version in 2003, depicts the health system as a way for service delivery, human resource planning, increasing access to healthcare as well as improving the standard and quality of care. According to the Health Systems Review of 2013/2014, the Department of Health (DoH) focused on strengthening the systems and improving capacity with the main goal to improve the quality of care and making the public health system more sustainable. The Negotiated Service Delivery Agreement was also developed to improve effectiveness by re-engineering the primary healthcare, human resources and standard-

[1] The OECD was formed in 1960 and comprised of 20 countries which have since grown to 34, including countries such as New Zealand, Korea, Japan, Germany, France, Sweden, Switzerland, Canada and Finland (Ngoepe, 2016).

ising the quality of care by certifying health institutions. The initiatives in South Africa make use of international support from institutions such as the World Health Organisation, the World Bank, bilateral donors, and private donors (Health Systems Trust, 2014).

In order to attain the goal of standardising, measuring and enforcing quality of care throughout the country, the National Core Standards (NCS) were developed starting from 2009 and implemented from 2011. The NCS take on a systems approach and provides benchmarks across seven overlapping domains, namely:

1. The rights of patients;
2. Patient safety, care and clinical governance;
3. Clinical support services;
4. Public health;
5. Leadership and corporate management;
6. Operational management; and
7. Infrastructure (Health Systems Trust, 2014).

A subset of non-negotiable standards was also set out with a focus on patient concern inside the seven domains, which are:

1. The attitude and values of staff;
2. Decreasing queues and waiting time;
3. The cleanliness of the facilities;
4. The safety of the patients and reliability of care;
5. Prevention of transmitting infections in the facilities; and
6. The availability of supplies and equipment (Health Systems Trust, 2014).

Despite the various initiatives and investments, South Africa's health system has been and is currently battling with key challenges including:

1. Multitude and complex diseases;
2. The quality of public healthcare;
3. An inefficient and ineffective health system; and
4. Increasing costs of private healthcare (Department of Health, 2014).

The National Development Plan (NDP) comprises nine goals for South Africa's health system to achieve by 2030. Five goals are in line with improving health and well-being whereas the other four aim to strengthen the health system. The nine goals of the NDP strive to:

1. Achieve a minimum life expectancy of 70 years;
2. Work systematically to improve methods to prevent and cure tuberculosis;
3. Reduce infant, child and maternal mortality;
4. Reduce occurrence of non-communicable diseases;
5. Halve the accidents, violence and injury levels from 2010;
6. Complete the Health System Reforms[2];
7. Provide care to communities through primary healthcare teams;
8. Provide universal health coverage; and
9. Employ skilled and competent staff (Department of Health, 2014).

The NDP also highlights nine priorities required to achieve a more effective health system, for example improving the information systems, human resources, public-private partnerships and reviewing management positions (Department of Health, 2014).

2.1.3 Progress towards the Health Millennium Development Goals (MDGs)

In September 2000, the United Nations Millennium Declaration was signed by 191 of the member states to fight poverty, disease, hunger, illiteracy, damage to the environment and female discrimination. The MDG each with specific targets and indicators, were developed from this and comprise eight goals which should have been achieved by 2015, using 1990 as a baseline (Barron & Pillay, 2014). The eight interdependent goals were (World Health Organization):

1. Eliminating severe hunger and poverty;
2. Achieving comprehensive primary education;
3. Improving gender equality;
4. Reducing child mortality;
5. Improving maternal health;
6. Fighting diseases such as the human immunodeficiency virus (HIV) and malaria;
7. Working towards sustaining the environment; and
8. Developing global partnerships.

[2]National plan to reform the healthcare sector, mainly focussing on financing.

In 2009, the Lancet released a series of papers on South Africa's healthcare system, one specifically relating to South Africa's progress to achieve the Millennium Development Goals. Since the end of apartheid up to 2009, the South African government focused on reducing inequality in the health sector. Primary health services for maternal care and childcare were supplied free of charge, 1300 clinics were built and abortion was legalised. In these 15 years, with the new health policies and changes, it was reported that insufficient (goals 2 and 6) or reversed progress (goals 1, 4 and 5) was made towards achieving the MDGs (Chopra *et al.*, 2009). The life expectancy at birth has decreased with almost 20 years since 1994. This resulted in a life expectancy of 50 and 54 respectively for men and women. The total disability-adjusted life-years for high-burden diseases in South Africa was almost the same as that for Bangladesh, which has a population about three times that of South Africa who lived in poverty. It was also reported that there was a high increase of people requiring chronic care for mental illness, HIV and tuberculosis (TB), and non-communicable diseases. This lead to the Lancet proposing that the government re-examine the resource-distribution between the various areas of care (Chopra *et al.*, 2009). The series identified four major health challenges which were (i) maternal, newborn and child death, (ii) HIV and tuberculosis, (iii) chronic diseases and mental health, and (iv) violence and injury.

By 2012, the picture described in the Lancet review of 2009 has changed, in some cases to be more positive. Since 2009, a ten-point plan was developed and the National Service Delivery Agreement was accepted. This plan focused on four goals, namely to (i) increase the life expectancy, (ii) decrease child and maternal mortality, (iii) decrease the HIV and tuberculosis disease burden, and, (iv) improve the effectiveness of the health system (Department of Health, 2014). This resulted in the life expectancy increasing from 2010 to 2012 to 60 years, and reducing child mortality and infant mortality respectively from 56 to 40 children per 1000, and, 40 to 30 infants per 1000. This is also regarded as potentially the biggest factor for preventing the transmission of HIV from mother to child (Motsoaledi, 2012). In 2009, South Africa implemented the pneumococcal conjugate vaccine[3] and rotavirus[4] vaccine in the Expanded Programme for Immunisation, which was the first of its type in Africa. By 2012, it was fully integrated into the programme. Since 2008 there has been a lower incidence rate of pneumonia and diarrhoea in children under five (Barron & Pillay, 2014).

From the Millennium Development Goals, further targets have been set. One is to eliminate malaria in South Africa by 2018, which means that there should be zero cases of malaria transmittance. Since 2000, there has been a 90% decrease in the number of cases and an 80% decrease in related deaths (Barron & Pillay, 2014). Since 2009 the annual number of tuberculosis cases and resulting deaths have decreased. Preventing HIV infected persons also contracting tuberculosis have since 2009 increased and the national programme for tuberculosis testing with the GeneXpert device, which is the largest in the word, was launched (Barron & Pillay, 2014). In 2012 South Africa's antiretroviral therapy programme was the largest in the world, with about 1.9 million people taking the triple dosage. Similar accomplishments have also been made in the treatment and testing of tuberculosis and HIV (Motsoaledi, 2012). Even though the health system is improving, there are still several challenges to be overcome, one of which is the poor accountability and effectiveness by management on all levels of the sector (Motsoaledi, 2012).

[3]A vaccine that can protect children and adults from pneumococcal disease which can cause ear, lung, blood and brain infections.

[4]Rotavirus infections are of the leading causes of severe diarrhoea among children.

2.2 Psychiatric care in South Africa

The WHO and various authors worldwide have produced various reports that increased the awareness of disability, the availability of cost-effective treatments and the economic costs associated with mental disorders. In May 2013, the comprehensive Mental Health Action Plan (MHAP) set out by the World Health Assembly, committed all United Nations member states to certain tasks that will enable them to reach a set of targets from 2013 to 2020 (Stein, 2014). The main objectives aim to:

1. Improve the effectiveness of leadership and governance of mental health;
2. Provide integrated, responsive and all-encompassing social care in communities;
3. Implement plans to promote mental health and prevention methods; and
4. Strengthen the information systems and research for mental health (Stein, 2014).

In South Africa, there has been a similar realisation with regard to mental health being neglected since before the democratic system was instituted. The Mental Health Act published in 2002 served as the first and significant improvement with regard to acknowledging the human rights of people suffering from mental disorders as well as their right to have access to care. Various provincial and national mental health conferences took place in 2012, which led to the National Health Council adopting the Mental Health Policy Framework (MHPF) for South Africa in 2013, along with the above-mentioned MHAP for 2013 to 2020, which includes mental health-related tasks and goals.

2.2.1 Mental illness

Mental illness or disorder refers to various health conditions that affect the behaviour, thinking and/or mood of a person. A mental health concern affects most people for short periods of time, but as soon as symptoms and signs become frequent and prolonged, affecting a person's ability to function and cause stress, it becomes known as a mental illness. In most cases, symptoms are treated with both medication and psychotherapy (Mayo Clinic, 2014).

There exist a variety of different mental disorders with varying severity that may be grouped in classes. The most common diagnosed disorders include depression and anxiety, panic attacks, obsessive-compulsive disorder, phobias, bipolar disorder and schizophrenia (South African Federation of Mental Health, 2011). Some of the more common classes are as follows:

Anxiety disorders result in affected people responding with fear to certain objects or situations and displaying physical signs of panic such as increased heartbeat and sweating. It is diagnosed when the response is seen as inappropriate in the situation, it interferes with normal functioning and one cannot control the response (Goldberg, 2014). Anxiety disorders include generalised anxiety disorder, panic, social anxiety and specific phobias.

Mood (affective) disorders involve recurrent and prolonged feelings of sadness or excessive happiness and fluctuations between the two. Depression, bipolar disorder and cyclothymic disorder fall in this category (Goldberg, 2014).

Psychotic disorders relates to a person's awareness and thinking being distorted (e.g. schizophrenia). Two common symptoms are hallucinations and delusion (Goldberg, 2014).

Eating disorders are related to adverse emotions and behaviour with regard to food and weight. The most common examples are binge eating, anorexia nervosa and bulimia nervosa (Goldberg, 2014).

Personality disorders may cause a person problems at work, school and/or relationships. A person's thinking and behaviour might differ from the general society and are usually extreme and rigid personality traits. Examples include obsessive-compulsive personality disorder, paranoid personality disorder and antisocial personality disorder (Goldberg, 2014).

Addiction and impulse control disorders results in the person suffering from the illness being unable to resist urges to take some sort of action that my harm themselves or others. Examples include pyromania (igniting fires), kleptomania (stealing) and uncontrolled gambling. Addiction involves a person that becomes so involved with an object such as drugs or alcohol that they become irresponsible and disregard relationships (Goldberg, 2014).

Post-traumatic stress disorder generally develops after a traumatic event, for example physical or sexual assault, a natural disaster or the sudden death of a loved one. Symptoms include a person having long-lasting memories and frightening thoughts of the event (Goldberg, 2014).

Other less common diseases are presented in Appendix A.1. Some diseases are occasionally grouped with mental illnesses because they are related to the brain and include various sleep-related problems as well as different forms of dementia such as Alzheimer's disease.

2.2.2 Mental illness in South Africa

A South African Stress and Health study reported that only one in four people suffering from a mental disorder has received treatment. This is in accordance with a study done by the WHO which reported that between 76 and 85 percent of people in low-income to middle-income countries with severe mental illnesses receive no treatment (Khumalo, 2012). Another study conducted in 2003 by the South African Stress and Health committee reported that a third of adults will suffer from some kind of mental illness in their lifetime. In the twelve months during which the study was conducted, 16.5% of the adults suffered from a mental illness (Chiumia & van Wyk, 2014).

Mental illness can and do occur with non-communicable diseases such as cancers, heart diseases and diabetes as well TB and HIV (Khumalo, 2012). Research have reported that more than 40% of HIV-infected people in South Africa have a diagnosable mental illness. A study conducted by the University of Cape Town reported that one in every three women from low-income settlements in and around Cape Town suffer from postnatal depression. In rural KwaZulu-Natal, it was reported that 41% of pregnant women are depressed. This is three times higher than in developed countries. The South African Depression and Anxiety Group claims that one in six South Africans suffer from depression, anxiety or substance abuse problems, which do not include more serious diseases such as schizophrenia (South African College of Applied

Psychology, 2013). Studies have also reported that mental illnesses, after HIV and other diseases, are the third-highest contributor to the burden of local diseases (Sorsdahl *et al.*, 2012). There exists proof of a strong link between mental disorders and poverty as it is influenced by factors such as malnutrition (Sorsdahl *et al.*, 2012).

In 2009 the first population-based, large-scale study was conducted on the 12-month and lifetime prevalence of common mental disorders in South Africa. The lifetime prevalence for any disorder was reported to be 30.3% with 11.2% of people having two or more disorders and 3.5% three or more disorders. The most prevalent classes of lifetime disorders were anxiety- (15.8%), substance abuse- (13.3%) and mood (9.8%) disorders. With regard to individual lifetime disorders, alcohol abuse, depression and agoraphobia[5] (without panic) were the most prevailing disorders (Herman *et al.*, 2009). The Western Cape had the highest lifetime prevalence rate, followed by the Free State, whereafter the lowest rates where reported in the Eastern Cape and Northern Cape. With regard to 12-month prevalent disorders, anxiety disorders were the most common followed by substance use and mood disorders.

The study also reported that females suffered more severely from 12-month mental disorders than males (Herman *et al.*, 2009). People with a high income had an increased risk of disease compared to low-income to average-income groups. Divorced, widowed or separated individuals also had an increased risk and severity for any disease compared to married individuals (Herman *et al.*, 2009). The age group of 35 to 49 had a higher prevalence rate than the age groups of 18 to 34 years and 65 years and older.

Substance disorders were more common in men while mood and anxiety were reported more in women. Mood and impulse control disorders were reported to be more prevalent in younger individuals with substance use being the highest in the 35–49 age group (Herman *et al.*, 2009). Women were reported to have a higher risk for developing anxiety or depression disorders and men have an increased risk for substance use (Department of Health, 2013).

2.2.3 Healthcare facilities in South Africa

In this section, the (i) public healthcare facilities in South Africa, and (ii) the types of psychiatric institutions in South Africa are introduced. Appendix A.2 contains a more comprehensive description of each facility that is mentioned in this section.

The basic point of entry to health services in South Africa is at primary level by means of local clinics and community health centres where 'ambulatory patients', people who can walk and do not require a bed, are treated. The public healthcare facilities can be grouped as follows:

1. Primary healthcare facilities;
2. Secondary healthcare facilities; and
3. Hospitals grouped in:
   (a) Level 1 or district hospitals;
   (b) Level 2 or regional hospitals; and

[5] Agoraphobia is the fear of open or public spaces.

(c) Level 3 or tertiary hospitals, again categorised in:

i. Tertiary 1 or provincial tertiary hospitals;
ii. Tertiary 2 or national referral hospitals;
iii. Tertiary 3 or central referral hospitals; and
iv. Specialised hospitals.

Appendix A.2.1 contains a detailed description of each of the facilities mentioned with regard to the type of services that can be supplied as well as the personnel present at the facility.

In September 2007, the WHO published a report on mental health in South Africa. In this report, account was given on the country's mental health services. Throughout this report, the fact that data are incomplete in some provinces or age groups was stressed. Only the data that are accurate will be presented in this section. The various mental healthcare facilities are as follows:

1. Outpatient facilities;
2. Day treatment facilities;
3. Community-based inpatient units;
4. Community residential facilities;
5. Mental health hospitals; and
6. Forensic and other facilities.

These facilities are described in Appendix A.2.2. The total amount of facilities in South Africa is reported along with demographic information, where available.

The majority of psychiatric patients are treated in outpatient facilities and mental health hospitals, and psychotropic drugs are also mostly available in the hospitals and inpatient units. From the report of the WHO, it is clear that the DoH rarely keeps data on the gender of the patients, the diagnoses and whether the patient is a child or an adolescent. If records are kept, they are not used for service planning, but held in individual case files. The data on the amount of facilities and beds per 100 000 population in the 2007 WHO report still corresponds when compared to data released in 2013 by the South African Federation for Mental Health (SAFMH) (South African Federation of Mental Health, 2013).

The national mental health authority, known as the "National Directorate: Mental Health and Substance Abuse", provides support to the government regarding mental health legislation and policies. The authority also provides support to the provincial authorities who are responsible for service planning and management and regular quality control. The resources and budgets allocated to each of the nine provinces vary greatly and services in provinces are organised in terms of catchment areas (WHO, 2007).

2.2.4 Legislation, policies and goals

The first mental health policy for South Africa was developed in 1997. It however did not follow the more recent adopted protocols and was not published for dissemination (WHO, 2007). In 2002, the Mental Healthcare Act No. 17 (MHCA) was published and fully implemented in December 2004 (Janse Van Rensburg, 2007). The MHCA of 2002 made an important contribution by emphasising the human rights of people with mental disorders and the access to care (Stein, 2014). Along with emphasising human rights, the main aims were to improve service by means of a primary healthcare approach, to improve community care and to protect the public's safety. The MHCA states the responsibility of the government with regard to providing adequate infrastructure and systems; defines the various categories of mental care users; explains the procedures required for each category; and lists the responsibilities and roles of the practitioners. The National Health Act No. 61 (2003) states that the Government is the main role player in providing mental healthcare by being responsible for the allocation of resources and infrastructure (Janse Van Rensburg, 2007).

After an important collaborative process involving provincial and national mental health gatherings in the beginning of 2012, another positive contribution was made towards improving psychiatric care in South Africa, when the National Health Council decided in July 2013 to adopt the MHPF as well as the strategic Plan for 2013–2020. There are eight key objectives for South Africa's MHCA plan, namely: "district-based mental health services and primary healthcare re-engineering; building institutional capacity; surveillance, research and innovation; building infrastructure and capacity of facilities; mental health technology, equipment and medicines; inter-sectoral collaboration; human resources for mental health; advocacy, mental health promotion and prevention of mental illness (Stein, 2014)."

In 2003, South Africa was one of less than 40% of countries worldwide who had mental health legislation that has been developed after 1990 (Ramlall, 2012). In addition, the country further improved its status with regard to mental healthcare by means of the MHPF and strategic plan. The SAFMH launched a three-year plan in 2013 to create public awareness about the need for more beds allocated to mentally ill patients and the need to improve services for people with mental illness and intellectual disability (Patel, 2014).

2.2.5 Mental health expenditure

In 2007, the percentage expenditure on mental health from the DoH was unknown on national level. Only three provinces could provide information on their mental health expenditure, with Mpumalanga spending 8%, the Northern Cape 1% and North West 5% of their health budget on mental health. Some provinces could not generate the data owing to the budget for mental health being combined, especially on primary care level, with the general health budget (WHO, 2007).

Mental health incurs serious and somewhat unexpected economic and social costs. The costs are directly related to providing healthcare, but also includes indirect costs that involve loss of income and employment owing to reduced productivity and impaired functioning (Department of Health, 2013). On the one hand, the indirect costs associated with mental illness is between two to six times more than its direct costs in developed countries, and still higher in developing

countries (Department of Health, 2013). The first national representative survey in South Africa about mental illness determined that the earnings lost by adults with severe mental disorders were R28.8 billion in 12 months. This amount was equal to 2.2% of the GDP in 2002. On the other hand, the direct spending on mental healthcare towards adults was about R472 million (Department of Health, 2013). In 2013, the South African Stress and Health study estimated the loss in earnings due to mental illness from 2003 being R40.6 billion, again compared to national mental health spending of about R 665.52 million (Bateman, 2014).

2.2.6 Mental healthcare in the Western Cape

The Western Cape is a very cosmopolitan province in South Africa with a unique demographic profile, covering an area of 129 462 square kilometres. There is one metropolitan municipality and five district municipalities. The province had a population of 5 553 957 people and 1420 893 households in 2011. Afrikaans is spoken by 55.3% of the province, Xhosa by 23.7% and English 19.3% (Health Systems Trust, 2012). As previously mentioned, the Western Cape has the highest lifetime and 12-month prevalence of mental diseases (Herman *et al.*, 2009). The burden of mental diseases in the Western Cape is intensified by an increase in substance abuse which further increases the demand for psychiatric inpatient services (Thomas *et al.*, 2015).

2.2.6.1 General health structure

In 2001, the Western Cape DoH started developing a strategic health plan for the province to be implemented by 2020. The strategic health plan is based on the 1995 Health Plan and the Comprehensive Service Plan of 2010 (Western Cape Government, 2011).

The health services of the Western Cape comprise various institutions providing service on all levels of care as displayed in Figure 2.1. Community-based services, primary healthcare (clinics and community health centres) and district hospitals form part of the District Health Services, i.e. level 1 service. Regional Hospitals provide level 2 or general specialist services where there are also specialised hospitals for psychiatry and tuberculosis to name a few. The health service platform of the Western Cape also includes central hospitals which provide general- as well as highly specialised care (level 3 services). Across these levels of service, there are specialised services which include Emergency Medical Services (emergency- and planned patient transport services) and Forensic Pathology Services which include conducting post mortem- and death scene investigations. There are also services such as pharmacies, laboratories and administration services offering a support function throughout the various levels of care in the framework (Western Cape Government, 2011).

There are four Psychiatric hospitals in the Western Cape, namely Stikland-, Alexandra-, Lentegeur- and Valkenberg Psychiatric Hospital. The province also has two sub-acute facilities, namely New Beginnings and William Slater. Table A.2 in Appendix A.3 summarises the challenges and status of psychiatric hospitals from what was planned for in 2010, the status quo in 2011 and then the envisaged plan for 2020 (Western Cape Government, 2011). The strategic plan focus on patients with an intellectual disability, the capacity of hospitals, standardising procedures and implementing the Mental Healthcare Act.

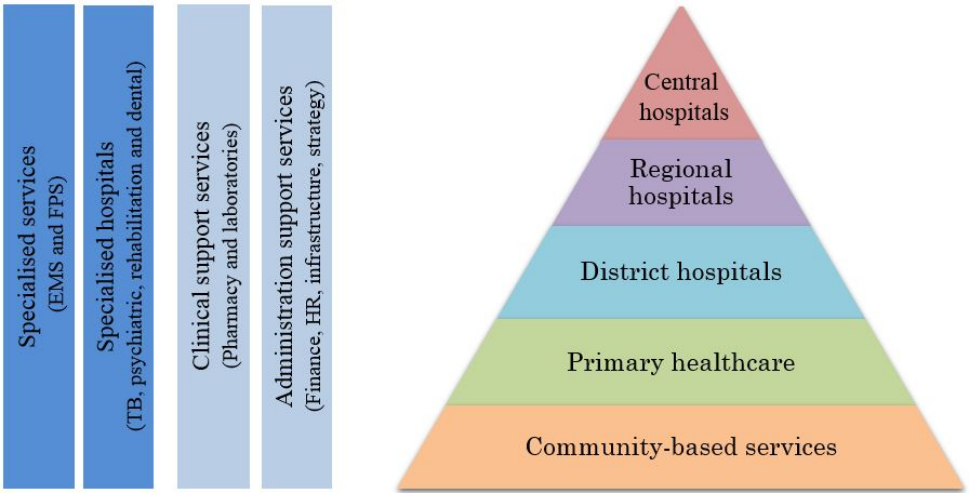


FIGURE 2.1: *Western Cape Healthcare Services platform. (Adapted from: Western Cape Government (2011).)*

2.2.6.2 General procedure for admitting psychiatric patients

In accordance with the MHCA 2002 no. 17, which encourages rehabilitative and community-based care, some public and regional hospitals were appointed to implement a 72-hour observational period on all mental health patients in an attempt to increase the accessibility to primary psychiatric healthcare. This involved admitting a patient (voluntary or involuntary), assessing the patient every 24 hours and treating them in order to exclude possible organic pathology[6]; allowing for the treatment of behavioural problems or possible substance or medically related psychiatric disturbances; and allowing for the patient to recover adequately in the period in order to attain good judgement with regard to giving consent to further treatment or be discharged. After 72 hours, two psychiatrists re-assess the patient and submit their recommendation to the head of the health facility who makes a decision with regard to further treatment at a psychiatric hospital (Thomas *et al.*, 2015).

The Western Cape strategic plan for healthcare in 2020, the pathway for service delivery to mental healthcare patients, is briefly described as an example for what is envisaged in 2020 as seen in Figure 2.2. At community-based service level, the patient receives care at home, along with regular visits from home-based carers to ensure that the patient is taking the prescribed medication, attending appointments or support groups and determine whether the family is understanding and supportive. In the case where a patient gets an acute psychotic episode, ambulance personnel will make the decision to either transport the patient to the nearest public healthcare clinic, community health centre or district hospital (Western Cape Government, 2011). At this facility, the patient will be subject to 72 hours observation by a competent doctor or nurse. If the patient does not respond to sedation or suffers from severe symptoms, they will be transferred to a psychiatric hospital, which may occur during the observation period. After a patient has been treated, a decision will be made to either discharge them or transfer them to an intermediate care facility. Throughout this process, various referral, admission and discharge

[6] An organic mental disorder describes decreased mental function owing to a physical disease.

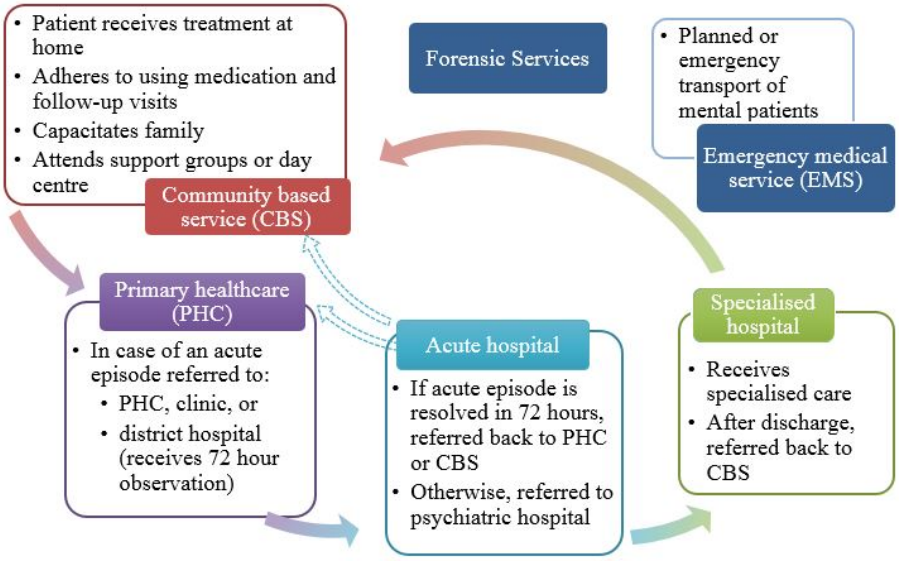


FIGURE 2.2: *Example of the path envisaged for a mental health patient throughout the various services in the Western Cape. (Adapted from: Western Cape Government (2011).)*

forms are completed. The patient will be encouraged to join a psycho-social group work and day-care centres which will assist in the recovery process (Western Cape Government, 2011).

2.2.6.3 Characteristics of the psychiatric population

A study conducted at Helderberg Hospital, which is a district-level public hospital in the Western Cape, brought to light information on the length of stay and general profile of psychiatric patients. Data from 2011 was used and included all admissions ranging from ages between 12 and 86, with a mean age of 34. Of the patients, 55% were male, and 51% of the admissions had a co-morbid diagnosis, with 42% relating to substance abuse, which, in turn, constituted 57% of the readmissions of which 65% were readmitted within three months (Thomas *et al.*, 2015). Diagnoses related to psychotic diseases were the most prevalent (59%). Within the psychotic class, mental health issues relating to substance use (56%) and schizophrenia (38%) were the highest. Substance use disorders followed psychotic diseases, with methamphetamine (commonly known as tik), alcohol and cannabis being the most abused substances. The third most common clinical diagnostic class was mood disorders with depression ranking the highest (58%) (Thomas *et al.*, 2015).

More men than woman was diagnosed with substance induced mental diseases, especially the abuse of methamphetamine (Thomas *et al.*, 2015). On the other hand, females where overly diagnosed more with mood disorders. Psychiatric and methamphetamine related admissions where the highest inpatients between the ages of 21 and 30. Of the admissions, 18% were transferred to tertiary psychiatric hospitals, while the majority were discharged after receiving treatment at Helderberg Hospital (Thomas *et al.*, 2015).

The mean length of stay of patients who had to be transferred to a psychiatric hospital was 9.4 days while those who stayed at Helderberg Hospital only stayed 4.7 days. The average number

of days a patient stayed at Helderberg Hospital before being transferred to another hospital, was 14 days for Tygerberg, 10 days for Stikland, six days for Alexandra and nine days for Lentegeur (Thomas *et al.*, 2015). The queuing time (day of admission to transfer) varied widely between the hospitals and the long waiting times may be explained by limited beds for certain population groups, e.g. children (Tygerberg Hospital) and the elderly (Stikland). Dementia cases had the longest length of stay followed by substance abuse-related cases. The study reported a definite association between length of stay and age, but not with gender (Thomas *et al.*, 2015).

A similar study conducted in the Northern Cape also reported that the majority of patients are male (68%) with a mean age of 32, diagnosed mainly with schizophrenia and substance-related disorders (Thomas *et al.*, 2015). A different study conducted in Cape Town psychiatric hospitals also reported that most patients are male, with a mean age of 25 and diagnosed with substance-induced psychosis followed by schizophrenia. Methamphetamine have been proven to be associated with mental disorders such as depression, suicide and psychosis. It is also the most abused substance in the Western Cape by men with a mean age of 26 years (Thomas *et al.*, 2015).

2.2.7 Stikland Psychiatric Hospital

In this section, a brief introduction is given on Stikland Psychiatric Hospital whereafter the processes with regard to the acute male ward is described. Affiliated outpatient care programmes are also introduced.

2.2.7.1 The hospital

Since April 1962, when Stikland Psychiatric Hospital was officially opened, it has been delivering secondary and tertiary level mental healthcare to the rural region of the Western Cape and the Cape Winelands as well as Eastern Metropole districts, Panorama North and Tygerberg. In 2001, the South African census revealed that the hospital serves a population of 1426791 citizens (Stikland Hospital).

Stikland Psychiatric Hospital, further referred to as Stikland Hospital, offers various mental healthcare services. Inpatient services consists of various short and medium term wards and includes a psychotherapeutic ward, psychogeriatric wards (elderly patients) serving the whole of Western Cape, acute male and female wards, an alcohol rehabilitation ward and an opioid detoxification ward (Stikland Hospital). The acute wards serves patients from the Stikland Hospital drainage area, and these patients generally have psychotic or mood disorders. Alcohol rehabilitation are for voluntary patients and also serves the whole Western Cape and the programmes are three to four weeks long. Outpatient services are referral based (from any health care practitioner), for new evaluations and follow-up appointments. Other services include psychology, occupational therapy and social services (Smit, 2016).

It is required that patients are first evaluated by their private doctors or day hospitals after which they may be referred to Stikland Hospital. In the case where a patient is resistant to an evaluation or unmanageable by a private doctor or day hospital, the closest police station to the patient is usually contacted for assistance. This is in accordance with the procedures as stipulated in the Mental Healthcare Act no. 17 of 2002 (Stikland Hospital; Smit, 2016).

2.2.7.2 General practices and operating procedures specific to the research

An acute male patient typically arrives at Stikland Hospital as a referral from a district hospital or community healthcare centre. In most cases, the patient has undergone the prescribed 72-hour observation prior to this, but, if not, the patient will be subject to a 72-hour observation at Stikland Hospital. The patient will be assessed by a doctor whereafter they will be admitted to the acute male wards. Stikland Hospital works in close conjunction with institutions and community-based treatment programmes that provide support to patients who are discharged. Two such institutions are the Assertive Community Treatment (ACT) team and New Beginnings (Stikland Hospital).

At present when a patient has to be crisis-discharged in order for a new patient to be admitted, the first thing the ward doctor does is to consult the nursing staff who will supply the psychiatrist with three or four names of patients who are not a management problem in the wards (Smit, 2016). When a patient is to be moved to New Beginnings, but there is not a bed available yet, the patient will be kept at Stikland Hospital for the time being, if possible. Sedation is also an indicator that is considered. Patients who have not been receiving sedation for a longer period of time will be discharged in stead of a patient recently taken off sedation or still on sedation (Smit, 2016). The home-situation of a patient is also considered. A patient with a more supportive environment will rather be discharged than a similar patient who stays for example, at a shelter. ACT patients are also more likely to be chosen for discharge than similar non-ACT patients. Schizophrenia patients are also seemingly discharged quicker than those suffering from a mood disorder. The rationale behind this is that a patient with a mood disorder can become more uncontained at home than a patient with psychosis and thus will have a higher chance of requiring readmission directly after discharge (Smit, 2016).

Most patients who are discharged from the acute male wards at Stikland Hospital are considered crisis-discharges according to the provincial definition (Smit, 2016). Patients who are not often selected to be crisis-discharged are for example, patients who are admitted under the influence of copious amounts of alcohol or illicit drugs. They usually improve in a matter of days and are discharged on no medication. Further admission to an acute ward would be of no benefit to these patients (Smit, 2016). Suicidal patients are also admitted and then discharged after about three days when their symptoms have improved (Smit, 2016).

There is a waiting list for patients requiring admission from various referring hospitals. The acute male wards have about 60 to 70 admissions per month. In effect the same amount of patients needs to be discharged to accommodate those being admitted (Smit, 2016). The system is also complicated by the concept around direct admissions, which occurs when a clinic refers a patient from an area in the Stikland drainage area that does not have a level 1 hospital. Stikland Hospital cannot refuse direct admissions and therefore these patients are prioritised for admission ahead of those already on the waiting list (Smit, 2016). The MHCA stipulates that Stikland Hospital must accept a patient for admission within 48 hours from the time the patient has completed the 72-hour observation at another hospital and require more time at a psychiatric institution. However, given the high turn-over in the wards needed to accomplish this time frame, it is not always possible (Smit, 2016).

The areas that do not have level 1 hospitals and results in direct admissions to Stikland Hospital, include Belhar, Elsies River, Ravensmead, Ruiterwacht, Bellville South, and Bishop Lavis (Smit,

2016). Patients from Eerste River, Vredenburg and Karl Bremer Hospitals are placed on the Stikland Hospital waiting list, and accepted for admission as soon as possible (Smit, 2016). All patients, even from areas that will require direct admission, must first be seen by a health care provider in their area and assessed before referral. However, in very rare occasions, patients can be admitted without prior evaluation, for instance a behaviourally disturbed patient who is already at Stikland outpatient services (Smit, 2016). Patients that are admitted under the MHCA as involuntary patients and must be evaluated by two independent medical practitioners (Smit, 2016).

Admissions for acute males at Stikland Hospital occur in ward five and discharges from ward four and eight. The management of all patients falls under the supervision of one psychiatrist, assisted by four general doctors (Smit, 2016). Weekly ward rounds are conducted by the psychiatrist whereafter all the patients' details, treatment and progress are discussed with the general doctors of that unit. Treatment is administered by the nurses who also have regular interaction with the patients (Smit, 2016).

2.2.7.3 The ACT team

Providing adequate community-based care relies especially on the availability of resources in the community. Community-based psychiatric services include clinics, day centres, group homes, home-based care along with adequately trained staff (Botha *et al.*, 2008). Globally, there are difficulties with setting up these services. Some countries, especially South Africa, have a unique combination of aspects complicating the implementation of community-based care. For example, the development of adequate resources for community care did not correspond to the decrease in the number of inpatient beds and in combination with poor socio-economic circumstances of the patients, it also contributed to the occurrence of the 'revolving door phenomena' (Botha *et al.*, 2008). Another obstacle in South Africa is the scarcity of day and residential services. Many families of patients also live with socio-economic difficulties and although patients with severe and long-term mental illnesses receive a disability grant, it is in most cases not adequate to reduce the financial pressure on the family. In addition, for some households, it is the only form of regular income. Healthcare staff, in turn, must cope with increased pressure on beds owing to deinstitutionalisation and crisis-discharge policies (Botha *et al.*, 2008).

Assertive community treatment (ACT) started in the 1980s and is focused on improving the quality of patients' lives by following a multi-disciplinary approach, being community-based, and providing intensive, frequent and comprehensive support to the patient (Botha *et al.*, 2008). In January 2007, ACT, an Associated Psychiatric Hospitals-driven programme focusing on providing support to high frequency users, generally uncooperative and who do not attend clinic appointments consistently, was introduced at the three psychiatric hospitals in the Western Cape (Stikland, Valkenberg and Lentegeur) (Botha *et al.*, 2008). Patients who have little to no support from family, are substance abusers or are not compliant to taking medication, make excellent candidates for admission to the ACT programme (Smit, 2016). The ACT is based at Stikland Hospital and the team comprises a senior social worker, chief professional nurse and a principal medical officer. ACT is a follow-up programme for acute patients with regular admissions, aiming to be more comprehensive than standard care (Botha *et al.*, 2008).

Generally, candidates for the programme are identified at admission and correspondence with

the family and patient is initialised during the inpatient stay. After discharge, the patient is followed up actively according to the patient's needs. On average, a patient is contacted at least once every two weeks – either by visiting the patient's home or at the psychiatric hospital or community mental health clinic, where medication is also dispensed. Most contacts are executed by the social worker or nurse with the medical officer contacting the patient monthly for the first three months, changing to once every three months after the patient is stable. Between visits, the patient is phoned and office phone numbers are also provided to the family. In the case of an after-hours crisis, ACT patients will be immediately assessed at the after-hours facility at the psychiatric hospital (Botha *et al.*, 2008).

The majority of high frequency patients (revolving door patients) access public healthcare services, live in adverse social environments and are unemployed, receiving disability grants. Some live in overcrowded and chaotic environments with financial difficulties, with no money for regular meals or travelling to clinics (Botha *et al.*, 2008). Patients also do not always have telephone numbers and in some cases home visits cannot be conducted when the safety of the healthcare worker is a concern due to gangsterism (Botha *et al.*, 2008).

In 2010, the ACT programme's performance was evaluated, which was the first study of its kind in a developing country. Previous studies conducted in developed countries reported that assertive intervention was not more advantageous for patients than standard care. Standard care in those countries may be understood as being much more comprehensive than in developing countries and in recent years have been incorporating some of the aspects of assertive treatment plans (Botha *et al.*, 2010). The study conducted at Stikland Hospital, however, reported that in a developing country, where standard mental services generally lack in resources, modified assertive treatment services do produce significant results such as a reduction in the amount of inpatient days and readmissions. Furthermore, assertive treatment services have a positive effect on the level of functioning of a patient (Botha *et al.*, 2010).

2.2.7.4 New Beginnings institution

New Beginnings was established in 2008 and is a residentially based rehabilitation institution for patients with severe mental illnesses. The facility has 40 beds with a multi-disciplinary team that assists with stabilising discharged patients who still require some kind of transitional supervision. The residents at New Beginnings follow a diverse programme that addresses substance abuse, work and life skills and relaxation therapy. Some also have the choice to attend weekly sessions at two other rehabilitation programmes, namely Chris Steytler and Fountain House. The residents are also involved in activities pertaining to managing the facility, such as reception work, maintenance, working in a tuck shop as well as producing products such as beaded bookmarks, soap and cards (Stikland Hospital).

Initially, only patients from Stikland Hospital, could be referred to New Beginnings, which would then be referred to as a 'step down' facility (Smit, 2016). It is, however, also a 'step up' facility with regard to ACT patients also being allowed at New Beginnings, or patients that have previously been at New beginnings, being able to be referred from home. Going to New Beginnings is voluntary, whereas at Stikland Hospital the patient can be kept against his will under the MHCA. On average, people stay about three weeks, although the ideal would be to complete the three month programme (Smit, 2016).

New Beginning Residential, which is under the same management, is a programme offering long-term stay and is a place of permanent residence for those patients. In the case of this project, the majority of the patients are at the 'step down, step up facility' (Smit, 2016).

2.2.7.5 Co-morbid substance abuse among psychiatric patients

In 2008, a study was conducted at Stikland Hospital to investigate the prevalence of substance abuse disorders in acute adult inpatients. The study reported that 51% of the patients had co-morbid substance abuse or dependence (Weich & Pienaar, 2009). This was reported especially among patients who are younger; involuntarily admitted; display above average violence before admission; and/or prefers using methamphetamine or cannabis. The conclusion was made that substance abuse disorders are prevalent and do contribute to the cognitive function of psychiatric patients (Weich & Pienaar, 2009).

In general, co-morbid substance abuse disorders contribute to various other aspects, such as patients being more likely to have behavioural problems such as aggression, non-compliance to treatment, readmissions, HIV-infection, carer distress, homelessness and becoming suicidal. It has also been reported that compared to patients with only one disorder, patients with co-morbidity may have increased treatment costs and suffer greatly from social isolation, unemployment and financial problems (Weich & Pienaar, 2009).

The study at Stikland Hospital reported that 72% of the patients with co-morbid substance abuse disorder were male. The highest substance abuse rates were reported to be inpatients from Belhar, Bishop Lavis, Elsies river, Kraaifontein, Paarl and Vredendal areas in the Western Cape (Weich & Pienaar, 2009). Schizo-affective patients and those with substance-induced disorders had the strongest association with substance abuse disorder. Only 7% of the patients were diagnosed with substance induced psychiatric disorder. This observation was deemed to be unexpectedly low and may be explained if the patients have been incorrectly diagnosed with schizophrenia, owing to the symptoms being similar to methamphetamine-induced psychosis (Weich & Pienaar, 2009).

2.3 Deinstitutionalisation and readmission

In recent decades there has been a global trend towards deinstitutionalisation in psychiatric care which led to a reduction in the length of hospital stay as well as the number of available beds (Heggestad, 2001).

Deinstitutionalisation entails releasing patients; reducing admissions and readmission rates; and shortening stays in order to reduce the population size of mental institutions. Furthermore, it also focuses on improving the institutional processes to reduce or eradicate behaviour that supports dependency, learned hopelessness and other abnormal behaviour (Stroman, 2003). In short, deinstitutionalisation leads to a faster transition from a psychiatric institution to the community. Some patients have, however, suffered serious problems or experienced no improvement after discharge, which again led to an increase in the number of readmissions and emergency referrals (Barekatain *et al.*, 2013). Evidence in support of deinstitutionalisation was deducted from studies conducted in the 1970s which reported that readmission rates did not differ between

short-term and long-term stays. These studies also suggested that longer hospitalisation makes it more difficult for patients to adjust to the 'real-world' (Niehaus *et al.*, 2008). Unfortunately, reports on the consequences of reducing the length of stay have been ambiguous. It is true that shorter length of stay is only effective with adequate outpatient care and discharge planning (Niehaus *et al.*, 2008). Deinstitutionalisation is in some cases also referred to as 'utilisation management' and also applied as a cost-saving strategy (Wickizer & Lessler, 1998). Some studies have been conducted to determine whether treatment restrictions do or utilisation management does in fact affect the quality of care.

It has been reported that a patient is most vulnerable for readmission during the period immediately after discharge (Durbin *et al.*, 2007). This implies the importance of the transition phase from hospital into the community along with the need for well-planned discharge policies and practices that may influence a patient to best manage the transition phase (Durbin *et al.*, 2007).

A study done by conducting trials to compare planned short stay, standard care and long stay reported that the patients receiving planned short stays experience less to no re-admissions, no losses to follow-up and were more successfully discharged compared to the other two options (Johnstone & Zolese, 1999). The study also reported some indication that short stay patients did not have an increased probability to leave a hospital too early and had an increased chance of being employed. The conclusion can be reached that hospital care and length of stay are important for mental policies. In addition, the conclusion reached suggests that planned short stay does not lead to the 'revolving door' phenomenon for serious mental illnesses (Johnstone & Zolese, 1999).

2.3.1 Indicators for re-admittance

The readmission rate is a popular research outcome measured and reported in mental health research. Finding trends in discharges and readmission of patients in a specific facility could help with effectively managing patients as well as distributing psychiatric resources.

Recurring admissions in a short period of time (usually three to nine months) are generally considered as avoidable and possibly damaging to the well-being of the patient. Furthermore, recurring admissions also entail greater costs for the healthcare system (Moss *et al.*, 2014). Deinstitutionalisation and the decentralisation of mental services have become more popular, especially in western mental healthcare systems. These policies have resulted in shorter length of stays and maintaining higher occupancy levels. As a result, readmissions are considered as a possible quality indicator for inpatient services (Moss *et al.*, 2014).

When a patient cannot be safely taken care of as an outpatient or in the community, they are transferred to inpatient care. Inpatient care is generally more expensive, more resource intensive and may pose more of an increased risk to hospital staff than outpatient care (Yussuf *et al.*, 2008). There are various reasons and variables that may contribute to a patient being readmitted. These are discussed in more detail in this section.

Readmissions entails additional expenses and may be disruptive to the patient and their families (Agency for Healthcare Research and Quality, 2014). Some guidelines to decrease readmission of patients entail providing adequate inpatient care and effectively stabilising the patient; applying

an appropriate discharge plan; and providing support with regard to the transfer process of a patient from being an inpatient to being an outpatient. Outpatient care generally includes discharge services, follow-up visits, short-term case management and provision of information to the patient and families with regard to the patient's mental illness (Agency for Healthcare Research and Quality, 2014). In 1984, 45% of the admissions to psychiatric hospitals in South Africa were readmissions (Gillis *et al.*, 1986). Recent studies conducted in the USA reported that between 2003 and 2011 the hospitalisation rate for mental disorders increased more than any other type of hospitalisation such as surgery, injury and maternal reasons (Heslin *et al.*, 2015).

Various studies have been conducted with regard to investigating factors that may be related to readmission to psychiatric hospitals. A literature review was conducted by using mainly search terms with 'readmission to psychiatric (or mental) hospitals' to search Google Scholar, Google, Science Direct and Scopus. Twenty articles were found to be relevant to this research. Special interest was given to articles that mentioned the data analysis methods, are applicable to psychiatric care and were available open source or by access through the Stellenbosch University. The articles were summarised and information with regard to the objective, sample size, study period, statistical methods, patient information used, variables tested and final results were gathered.

The studies used patient information applicable to mainly one institution, with a few exceptions. The objective of all the studies involved investigating whether there is common indicator or patient variable that has a strong relationship with readmission. The information would then be used for discharge planning, planning the length of stay for a patient and identifying specific needs for follow-up, to name a few examples. The studies were conducted around the world, from teaching hospitals in Africa to large private companies accessing public healthcare data.

With regard to size and sample-type, the studies varied. The majority regarded all psychiatric inpatients from one psychiatric facility or ward, while a few focused on one diagnostic type such as depressed patients or only male inpatients. The samples sizes ranged from 100 patients to around 22 000, depending on the sample space, inference level and time period of the study. The period over which the studies were conducted ranged from one year to ten years, with about 75% of the reviewed studies having a study period of between two and five years. In some studies, only patients admitted during a specific year were followed up for another few years, disregarding patients admitted in the following years, whereas other studies, ranging over a longer period, studied all the admissions during that period.

The studies reported multiple factors that are linked to readmissions at psychiatric hospitals. Shorter length of stay has been found to be linked to readmission along with more serious psychotic disorders, substance abuse, the male gender, marital status, unemployment and involuntary admission (Moss *et al.*, 2014). Along with these, history of violence or criminal behaviour, demographic and socio-economic characteristics, and non-compliance to treatment also contributed to patients being readmitted (Haywood *et al.*, 1995). It is, however, invalid to generalise from all the studies owing to the samples studied being small and limited to one institution. Studies have also been conducted to determine whether readmission rates can be an indicator of the quality of care at a mental hospital. Table 2.1 summarises some of the literature found on studies done to find factors that play a role in readmission rates. The most indicative factor appears to be previous admissions.

TABLE 2.1: *Indicators of readmission from previous studies.*

	Objective	**Sample size**	**Period (time)**	**Results**
(Barekatain *et al.*, 2013)	Determine the readmission rate, the social, demographic and clinical characteristics of patients admitted and the factors related to readmission.	Cross-sectional study with 3 935 patients admitted to the Psychiatric ward of Isfahan University Hospital, Iran.	2004–2010 (seven years).	The number of readmissions was most significantly explained by psychiatric anxiety disorder, bipolar, depression, psychotic disorder and a history of self-reported previous admissions. Education had no correlation and being divorced was reported to be correlated with more amounts of readmission. The number of readmissions per patient using opium were 1.3 times higher than patients not using the drug.
(Bernardo & Forchuk, 2001)	Determine patient-related factors associated with readmission.	A random sample of 200 patients from a tertiary psychiatric hospital in Canada.	Index discharge date in 1991 with rehospitalisation investigated up to 1994 (four years).	A strong correlation was reported between readmission and a history of admission. A weak inverse correlation between the age at first diagnosis and age at readmission was observed. The only significant indicator for readmission is a history of previous admissions. The number of patients decreased with the number of readmissions along with the mean length of stay and time between admissions. The diagnosis was also reported to be linked to readmission, with schizophrenia being the primary diagnosis for 41% of the readmitted patients, followed by personality-, mood- and schizo-affective disorder. Of the patients, 88% were readmitted at least once in the three-year study period. Of the readmitted patients, 59% were male. Readmitted patients were reported to be slightly younger at first admission than those who were not readmitted. Other significant factors related to readmitted patients were secondary education (compared to primary education); being divorced; unemployment, working part time or receiving social assistance and a history of aggression.
(Byrne *et al.*, 2010)	Determine whether hospital outcomes can be used as a predictor of readmission along with identifying indicators of readmission.	Patients with depression admitted for the first time (n=478) and all inpatients (n=1177), regardless of diagnosis and number admissions were compared.	Consecutively admitted patients in a psychiatric hospital between 23 June 1998 and 31 October 2003 (+/- five years).	Patients experiencing relationship problems had an increased chance for readmission. Gender (female) as well as age (younger) was found to be predictors for readmission for both the depression and general samples. Socio-economic status was only a predictor over the five-year period for the depressed sample. For both samples during the six-month and five-year follow-up periods, length of stay was reported to be a predictor for readmission. The study also concluded that readmission may be linked to the success of the previous hospitalisation. The predictors varied over long- and short-term periods and thus it is important to follow-up on patients after discharge.

Indicators of readmission from previous studies (continued).

	Objective	**Sample size**	**Period**	**Results**
(Durbin *et al.*, 2007)	Determine whether readmission rates can be used as an indicator for the quality of inpatient psychiatric care.	Studies published between 1995 and 2006. 13 studies met the inclusion categories out of 455 primary studies.	Literature over a period of around ten years.	A history of readmission increased the chance for being readmitted which suggests that special care should be taken to break the 'revolving door' phenomena. Background demographics (gender, race, marital and education) of a patient was not found to affect the risk for early readmission, except for age, where younger patients experienced readmission more frequently. During hospitalisation, the most significant indicator for early readmission was a history of previous hospitalisation(s). Diagnoses such as bipolar, depression and schizophrenia were reported significant in four out of ten studies. Unstable patients with active symptoms were most likely to be readmitted. Some evidence was found that suggested preparing patients for discharge and adequately stabilising them can reduce the probability for early readmission.
(Gillis *et al.*, 1986)	Determine factors that influence readmissions at Valkenberg Hospital, Cape Town.	Sample of 132 white, 139 coloured and 135 black patients followed up on intervals of six months for two years.	February 1981 to July 1984 (three years).	No significant indicators emerged from the demographic, personal, social and family variables pertaining to readmission within one year. Substance abuse, leading to many admissions, was also not found to be a significant indicator of readmission. There was a correlation between living alone and a higher admission rate with white patients, but it is difficult to judge owing to 24% of whites living alone and only 0.7% of coloured and blacks living alone. Patients with schizophrenia and affective disorders had a greater risk for readmission within one year and especially significant with coloured and black patients. More acute psychotic diagnoses were observed in black and coloured samples, mostly explained by higher drug and alcohol abuse (especially cannabis). Younger persons were also associated with more readmissions, which may be explained by the group having more acute diseases, especially behavioural disturbances. Within the coloured sample, a history of previous admissions, male gender and schizophrenia were significant indicators for readmission within one year. Previous admissions, living alone, absence of emotional support and mixed substance abuse were indicators for readmission in white patients.
(Haywood *et al.*, 1995)	Examine the relationships between demographic features, diagnostic characteristics and the number of hospitalisations among patients with schizophrenic, affective disorders and schizoaffective conditions.	135 inpatients (86 male and 49 female) were interviewed at four hospital facilities in the state of Illinois' Department of Mental Health.	Unknown.	Non-compliance with treatment and substance abuse were found to be significantly related to a higher number of admissions. The study reported that more women were first-admission patients and men were more frequently hospitalised, with 64% of men experiencing five or more hospitalisations. Marital status was not reported to be related to the number of admissions, but was reported to be associated with other demographic features, e.g. persons not married were on average ten years younger than married patients and men were more likely to be single, thus the interrelationship between characteristics complicates the accuracy in finding a relationship between marital status and number of admissions. Drug and alcohol abuse were reported to be associated with more frequent readmissions and patients with a higher number of readmissions were more likely to also be non-compliant with treatment.

Indicators of readmission from previous studies (continued).

	Objective	**Sample size**	**Period**	**Results**
(Heggestad, 2001)	Investigate the relationship between operating conditions in psychiatric hospitals and the risk of early readmission.	5 520 patients from about 20 Norwegian psychiatric hospitals.	1992–1996.	The hazard ratios indicated that high patient turnover (annual discharges per bed) results in a higher risk for readmission. Discharges from wards with lower access to therapists increased the risk for readmission.
(Heslin *et al.*, 2015)	A briefing about the statistics regarding hospitalisations of mood and substance use disorders (M/SUD) and non-M/SUDs.	n.a.	2011.	Schizophrenia and mood disorders were the two most popular diagnoses. Patients with mood disorder were twice as likely to be readmitted within 30 days compared to non-M/SUD patients. Compared to non-M/SUD conditions, LOS for schizophrenia and mood disorders were about 39% and 50% longer and they were also more likely to be readmitted with the same diagnosis.
(Innes *et al.*, 2015)	Investigate the initial admission for major depressive disorder (MDD) as well as readmissions pertaining to MDD.	Data from the Scottish Health Survey as well as historical admission data of 52 990 adults who have not been hospitalised for a depression-related incidence.	1995–2011 (16 years) (median follow-up of 4.5 years per participant.)	During the period, 530 participants were admitted for MDD for the first time. 118 were readmitted again during the period of study. The only two factors associated with readmission were older age ($\geq$70 years) and previous admission for a psychiatric illness other than MDD.
(Johnstone & Zolese, 1999)	Investigate the effect the length of hospitalisation has on patients with serious mental illness.	Five randomised trails and literature studies.	Unknown.	Compared to standard care and long stay care, no increase in readmissions and losses to follow-up were observed in patients receiving planned short stay. Planned short stay patients also had more on-time discharges. Short stay patients are not more likely to leave the hospital early and have an increased chance of being employed. Planned short stay policies do not contribute to the revolving door phenomenon.

Indicators of readmission from previous studies (continued).

	Objective	**Sample size**	**Period**	**Results**
(Jones *et al.*, 2002)	Compare readmission rates for adjustment disorders with that of mood disorders.	5 067 patients admitted to Laureate Psychiatric Clinic and Hospital (Oklahoma).	1990–1999 (ten years).	Diagnosis was reported to be a significant predictor of readmission. Adjustment disorders had fewer readmissions and major recurrent depression resulted in significantly more readmissions.
(Loch, 2012)	Investigate the rehospitalisation rates regarding psychosis and bipolar disorder as well as the indicators for readmission.	169 patients with bipolar and psychotic disorder, from one of the five public psychiatric hospitals, Hospital Psiquitrico Philippe Pinel, with 36 acute psychiatric beds for males and 12 for females.	Patients discharged from May to August 2009 were followed up after discharge on one, two, six and 12 month intervals.	The study reported that the three main variables influencing the likeliness of readmission are adherence to medication; the severity of the diagnosis; and, family agreeing to, or requesting permanent hospitalisation. Of the patients, 42.6% were rehospitalised within one year. Being physically restraint during hospitalisation and not attending follow-up sessions increased the risk for readmission. The number of previous admissions was also reported to be related to readmission. Emphasised by the study was the fact that family who wish to keep a patient permanently hospitalised also contributes to the revolving door readmission phenomenon. The fact that family requesting permanent hospitalisation is one of the main indicators for readmission, may indicate unwillingness by the family to look after a mentally ill patient.
(Lyons *et al.*, 1997)	Examine indicators for readmission to determine whether readmission can be used as an indicator for the quality of inpatient psychiatric service.	255 patients consecutively admitted to any of seven psychiatric hospitals in the regional managed care programme.	Patients admitted from July 1994 and February 1995 (7 months).	Of the sample, 49.3% was male, 17.6% were readmitted within six months and 7.1% within 30 days from discharge. With the 30-day readmission, the only significant indicator was a self-care impairment. Patients who experience medical complications were also reported to be readmitted within 30 days. Severe symptoms and substance abuse problems were common in patients who were readmitted in within six months. Predicting one year readmission, indicators were reported to be self-care impairment and the persistence and severity of symptoms. No evidence was found that suggested readmission within both 30 days or six months are associated with poor hospital care or early discharge.
(Malesu)	Investigate factors that influence readmission of schizophrenics.	208 patients.	Unknown.	Of the total readmissions, 70% were schizophrenics and 62% were male schizophrenics. The strongest predictors for readmission were the male gender along with age, homelessness, unemployment and a lower socio-economic class. The study also reported an increase in criminal behaviour, drug abuse and poor compliance to treatment in the male schizophrenic sample.

Indicators of readmission from previous studies (continued).

	Objective	**Sample size**	**Period**	**Results**
(Mark *et al.*, 2006)	Investigate the rate of detoxification readmissions and the factors associated therewith.	22 054 patients from Medicaid, state mental health and substance abuse agency databases from three states (Delaware, Oklahoma, and Washington)	1996—1998 (three years).	Patients who received more than one substance abuse service within 30 days of initial detoxification are less likely to be readmitted and have a longer time before readmission. Of the sample, 27% was readmitted within one year of initial admission. Readmission following detoxification is common in the public sector. The time to second detoxification for female patients was 25% longer than males. The time to detoxification also varied by race, with the duration to readmission for African Americans being about 20% longer when compared to Caucasians, who again took 20% longer when compared to Hispanics. More Caucasians and Hispanics were also readmitted when compared to African Americans. Older age was also reported to be indicative of readmission, but only differed slightly among races.
(Mayoral *et al.*, 2012)	Investigate factors linked to readmission in a sample of adult schizophrenic psychiatric patients.	100 adult psychiatric patients (58% men) discharged consecutively from a short-stay unit of a university general hospital.	One year.	Of the patients, 61% were readmitted during the 12 months after discharge. Functional status at discharge, work status and number of previous admissions (>2) were significantly associated with readmission.
(Moss *et al.*, 2014)	Determine the indicators for the time to readmission in a general psychiatry inpatient unit.	758 patients.	April 2006 to October 2008 (two years).	Of the patients, 21% were readmitted within 180 days from discharge. Of the sample, 45.3% were male. Most of the variables were not significant indicators for readmission. Previous admissions were associated with patients admitted one to two times in two years being 15.6 times more likely for readmission. Those admitted more than three times were 24.2 times more likely to be readmitted. This hospital hands out 'passes' to inpatients which allow the patient to leave the hospital in order to determine how the patient will function in the community after discharge. It was reported that patients with passes are 3.5 times more likely to be readmitted.
(Niehaus *et al.*, 2008)	Examine the effect crisis-discharge policies (de-institutionalisation) and length of stay have on the readmission rates in a psychiatric hospital in South Africa.	438 male inpatients at Stikland Hospital, admitted during 2004.	January 2004 to August 2006 (two years and eight months).	A shorter length of stay (LOS) was reported to be correlated with a decreased readmission rate, but crisis-discharges had a much greater effect on readmission rates, with patients that are crisis-discharged having a much greater risk for readmission. It was reported that a shorter LOS resulted in a longer time before the next readmission, but it cannot be concluded that a crisis-discharge policy contributes to the revolving door effect.

Indicators of readmission from previous studies (continued).

	Objective	**Sample size**	**Period**	**Results**
(Wickizer & Lessler, 1998)	Determine whether treatment restrictions imposed by utilisation management affect the chance for readmission.	3 073 reviews were performed on 2443 (51.7% men) privately insured psychiatric patients.	5 years (1989-1993).	The most common disorders were alcohol abuse, recurring depression and single-event depression. An average of 22.4 days of inpatient treatment was requested, with an average of 15.5 days being allowed due to the utilisation management programme. Of the cases, 7.9% were readmitted within 60 days of initial admission. Patients with restricted LOS were reported to be more likely to be readmitted. A reduction of ten days from the requested LOS resulted in a patient being 37% more likely for readmission within 60 days compared to patients whose LOS was not restricted. For each day the requested LOS is reduced, the chance for readmission within 60 days increased by 3.1% (P = 0.004).
(Yussuf *et al.*, 2008)	Identify indicators for readmissions which can be used to provide comprehensive mental care; minimise readmissions and also reduce over-utilisation of the facilities and some pressure on health staff; and to create a baseline on readmission for future research in the north central zone of Nigeria.	A retrospective record review of all admissions and discharges to and from the psychiatric inpatient ward of University of Ilorin Teaching Hospital, Nigeria (502 out of 789 records were complete).	May 2000 to April 2005 (five years).	Of the cases, 41% were readmissions. Four significant predictors for readmission were found. From the socio-demographic variables, younger age (21–40 years) were significant and from the clinical variables, a longer LOS, schizophrenia and non-compliance with medication were found to be significantly associated with readmission. These factors were analysed with logistic regression to determine which could predict readmission independently and all, except non-compliance with medication, emerged as indicators. Multiple admissions were also reported to be predictive.

2.4 The real-world problem described

Deinstitutionalisation and shorter length of stay are not unfamiliar terms in the psychiatric world with the length of stay in psychiatric wards decreasing internationally over the past few decades. The reason for this is not pressure on beds, but to give a patient the best health benefit along with maintaining higher capacity rates (Niehaus *et al.*, 2008). In South Africa the pressure on beds is high, which led to the Western Cape implementing a crisis-discharge policy. This is different from planned short stay as the patients are discharged earlier than what is clinically ideal for the individual patient in order to allow people on the waiting list to be admitted, thus optimising the combined healthcare outcome (Niehaus *et al.*, 2008).

Stikland Hospital is one of four psychiatric hospitals in the Western Cape. The Western Cape has from 2002 to 2007 experienced a 21% reduction in the number of psychiatric beds available, contributing to a national decline of 7.7% (WHO, 2007). In 2009, a study found that the Western Cape had the highest lifetime prevalence rate of common mental disorders in South Africa (Herman *et al.*, 2009).

A study conducted at Stikland Hospital's acute male ward concluded that patients who were crisis-discharged had an increased risk to be readmitted, and in the long run, that crisis-discharge might not decrease the capacity issues, but worsen it. This 'revolving door phenomena' led to the healthcare professionals question whether there is a way to improve the discharge methodology in order to reduce the readmission rates (Niehaus *et al.*, 2008).

Various studies have been conducted on the 'revolving door phenomena' which may occur from early discharge policies, but the results are ambiguous and seem to be case-specific. It is believed that shorter stay may make it easier for the patient to adjust to the environment outside the hospital, but it was found that this is only effective if outpatient care and adequate discharge planning are implemented (Niehaus *et al.*, 2008).

Mental health incurs both economic and social costs, which may be surprisingly high when considering the indirect costs pertaining to loss of income which can be up to six times more than the direct costs of healthcare (Department of Health, 2013). It is thus imperative to treat mental illness. In December 2004, the Mental Health Act No. 17 was fully implemented followed by the MHPF adopted in 2013 and the Strategic Plan for 2013 to 2020 (Janse Van Rensburg, 2007).

Studying readmission rates and indicators for readmission is popular in healthcare as finding trends at a specific facility may help with effectively managing the patients and resources (Moss *et al.*, 2014). Majority of the studies reported that a history of previous admissions were the most indicative of readmission.

This research is applicable to the acute male ward where readmission rates and capacity restrictions are a concern to the psychiatrists. The data is provided by the management of Stikland Hospital and pertains to admissions between 2012 and 2014. Data mining methods and biostatistics will be researched to investigate methods with which readmission rates can be investigated. The most significant predictors and the possible interrelationships resulting in the increased chance of readmission will be discussed and implemented to establish a form of decision support in order to minimise the revolving door effect. Predictive models are also built and evaluated and the feasibility of a decision-support tool is discussed.

2.5 Conclusion: The real-world problem

This chapter broadly explored the South African healthcare sector, introducing the current state of healthcare, legislations, goals and achievements. The real-word problem which relates to psychiatric care in South Africa was explored by describing mental illness, the various facilities, legislation, expenditure, mental healthcare specific to the Western Cape and finally readmission and deinstitutionalisation in particular.

Chapter 3 will introduce the field of data mining and biostatistics. The methods applied by similar published studies are also presented and the methods that can be possibly used for this research are investigated.

CHAPTER 3

The science of learning from data

The real-world problem was introduced in Chapter 2 by describing the state of South Africa's healthcare sector, focussing on the psychiatric health sector, deinstitutionalisation and readmission.

The amount of data generated by healthcare transactions have become more complex and voluminous and accordingly this chapter provides an introduction to data mining with a focus on the more popular data mining techniques. Biostatistics and data mining methods that can be used to analyse the data pertaining to this project are also discussed.

3.1 Introduction

The science of learning plays a fundamental role in the fields of data mining, statistics and artificial intelligence and overlap with areas in fields such as engineering, finance and the industry (Hastie *et al.*, 2009). Examples of problems solved by learning techniques are:

- Predicting whether a patient admitted to a hospital due to a heart attack will suffer from a second, by analysing demographic, diet and clinical information for the patient;
- Predicting the stock price in six months' time based on a company's performance measures and economic data;
- Identifying numbers from a digitised image of a handwritten ZIP code; and
- Identifying the risk factors for prostate cancer from demographic and clinical variables (Hastie *et al.*, 2009).

The majority of the methods employed in data mining, such as regression, classification, clustering and visualisation, were developed in statistics and machine learning. Data analysis is given various names, in most cases depending on the person as well as the type of data that is analysed. For example, the analysis of medical data may be referred to as biostatistics, but also as merely statistics, data mining, machine learning or artificial intelligence. The methods

also overlap between the aforementioned fields, making it difficult to introduce the concept of 'learning from data'. For the purpose of this project, the term data mining will be mainly used. The literature study included consulting various articles and textbooks, with keywords mainly pertaining to data mining, healthcare and statistics, as well as consulting statistical experts.

3.1.1 Data mining in general

Data mining entails selecting, investigating and modelling large amounts of data to discover unknown patterns and relationships, predict trends and remove excess data (Jacob & Ramani, 2012). Data mining methods are often grouped by their function: (i) description and visualisation; (ii) classification and estimation (predictive modelling); and (iii) clustering and association. Large data sets with especially complex and non-linear relationships can most effectively be understood by visualisation and descriptive methods. Predictive modelling involves probably the most common applications of data mining. Classification is used to predict a categorical target variable, for example if a patient will die, or not, where estimation refers to predicting a metric variable such as the length of stay or resource utilisation. The objective of clustering is to group objects, such as patients, so that the objects belonging to a certain cluster are similar. Association is used for determining variables that go together (Koh & Tan, 2005). Additional descriptions of these methods are presented in Appendix B.1.

A data set usually comprises instances, or entries, which contain values for a number of variables known as attributes. There are generally two types of data. The first is where the data, referred to as labelled data, is used to predict the value of a specially selected attribute for instances not yet seen (Bramer, 2007). In data mining terms, this is referred to as predictive data mining, where the aim is to build models for tasks such as regression and classification and then determining the accuracy of the model by applying it to a new data set (Izenman, 2008). Predictive data mining is also known in machine learning terms as supervised learning.

The second type of data is unlabelled data, used in descriptive data mining, which differs extensively from supervised learning, as it does not have a designated attribute and the aim is only to gain as much information from the data as possible (Bramer, 2007). The techniques for investigating unlabelled data focus mainly on searching through large data sets to discover the existence of unexpected relationships, patterns, clusters and outliers. Descriptive data mining is also referred to as unsupervised learning (Izenman, 2008). The aim in predictive data mining (supervised learning) is mainly to study relationships between input and output variables where in descriptive data mining (unsupervised learning) the goal is to explore characteristics of only input variables. This project will incorporate mainly predictive data mining methods owing to researching the relationship between certain patient variables (independent variables) and readmission (dependent variable). Data mining can be seen as similar to, or as a step in, a larger process known as 'knowledge discovery in databases (KDD)' which is mainly composed of:

1. Selecting the data (variables and attributes) to be used;
2. Data cleaning which entails removing noise, identifying outliers and imputing missing data;
3. Initial processing of data (tracking time-dependent entries and deciding on transforming the data);

4. Selecting the data mining method;
5. Analysing the prepared data set using software; and
6. Interpreting the results (Izenman, 2008).

3.1.2 Data mining methodology

In the industry, data mining projects are complex and may require combined input from various departments or stakeholders in the organisation. When data mining was still relatively new, scientists followed their own procedures and methodologies to complete their specific tasks. Various general methods, such as the previous mentioned KDD process, have been proposed to serve as guides for organising the process of gathering and analysing data, interpreting and applying the results and monitoring the progress (Statistica, 2015b). The need for a cross-industry standard, neutral to the application field, tools employed and industry to which it is applicable, becomes apparent. This led to analysts from five European companies, including Daimler Chrysler and SPPS, collaborating to develop the Cross Industry Standard Process for Data Mining (CRISP-DM) in 1996 (Larose, 2005).

According to CRISP-DM the life cycle of a data mining project consists of six phases, illustrated in Figure 3.1. The sequence of the phases may change according to the outcomes of the previous phase (Larose, 2005). According to a survey conducted in an online data mining community, CRISP-DM is still the most used methodology for data mining projects, and, although the analytical process is still applicable, it was mentioned that details and specifics might require updates to adapt to modern data science challenges (Piatetsky, 2014).

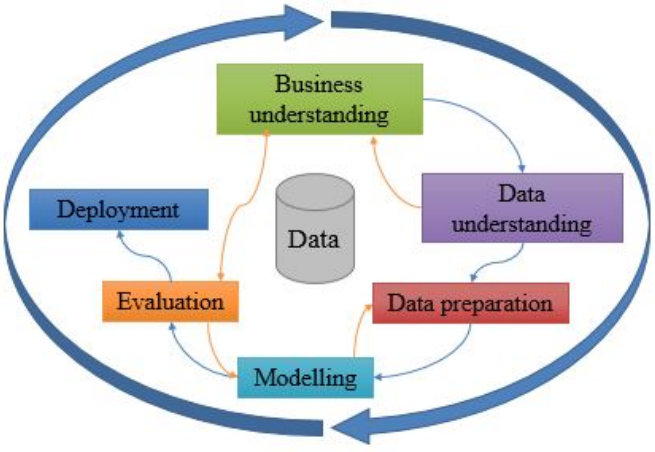


FIGURE 3.1: *The life cycle of a CRISP-DM project. (Adapted from: Larose (2005).)*

The previously mentioned survey found that after CRISP-DM, 35% of the data scientists commented that they use their own methodology (Piatetsky, 2014). A third methodology is SEMMA (Sample, Explore, Modify, Model and Assess), developed by the SAS Institute, which differs from CRISP-DM's comprehensive project management model by mainly focussing on the application to investigative visualisation and statistics-based data mining methods (Bellazzi & Zupan, 2008). Six Sigma is another data-driven approach for eliminating errors, waste and quality issues in

various business activities, especially manufacturing. The process steps can be applied to data mining and are as follows: define, measure, analyse, improve and control (Statistica, 2015b). All of the methodologies focus on how to integrate data mining with an organisation and how to convert data into information for stakeholders as well as into strategic decision-making. Most data mining software is designed to incorporate some sort of data mining framework (Statistica, 2015b).

3.1.3 Prediction accuracy and generalisation

The accuracy of a predictor, known as the prediction error, has to be determined. With a classification problem, the prediction error is described as the probability of misclassifying a case, where in regression, it is described as the mean of the squared errors. The regression error is the difference between the true output value and the predicted output value (Izenman, 2008). One of the simplest ways to calculate the prediction error is the resubstitution error. With classification, the classifier is evaluated by scoring a correct prediction as '0' and a misclassification as '1'. The proportion of misclassified cases is then viewed as the resubstitution estimate. In the case of regression, the fitted model is used to predict each of the known output values from the data set. The resubstitution estimate is calculated as the mean of the squared residuals. This method uses the same data to calculate the predictor and is thus not the most accurate method (Izenman, 2008).

The concept of generalisation was developed to improve the accuracy of the resubstitution estimator. This led to a need for a method that still makes good predictions when applied to an independent data set. An independent data set refers to the process of having to gather new data. This is not always feasible and leads to the practice of using a random procedure to split the initial data set if it is large enough, into three non-overlapping sets which are assumed to be generated by the same underlying distribution, namely:

1. A learning set ($\mathcal{L}$), which is used for preliminary testing, pattern searches, elimination of outliers and so forth;
2. A validation set ($\mathcal{V}$), which is used to select and assess competing models;
3. A test set ($\mathcal{T}$), which will be used for assessing the performance of the final model (Izenman, 2008).

The learning set can be taken from historical data or otherwise the data is split where 50% is used for learning and 25% is used for both validation and testing. The validation and test set is also sometimes combined to form a larger test set.

With supervised learning, the model (function of inputs) must be assessed in terms of how closely it fits the data (output). There are two types of prediction errors related to regression models. First a regression model must be fitted to the learning set, whereafter the fitted model is used to predict output values of either the learning or test set. The prediction error is the average of the squared errors in the predicted values from either the learning or test set. In the case where the learning set is used, the prediction error is called the regression learning error, and in the second case, it is known as the test error (Izenman, 2008). Similarly, in a classification

problem, the classifier is built from the learning set, whereafter it is used to predict the class of each vector in either the learning- or validation set. The prediction error is calculated as the proportion of misclassified observations (average of *0*'s and *1*'s divided by either the learning set or test set) (Izenman, 2008).

In some cases, it may feel that by only using a portion of the entire data set to fit a model, data is wasted, alternative data-splitting methods such as bootstrap and cross-validation can be explored to estimate test errors. V-fold cross-validation randomly divides the complete data set into v-overlapping, approximately equal sized groups, takes one group for the test set and v-1 groups as the learning set. The learning set is used to predict the output values after which the prediction error is calculated from the omitted group. The process is repeated v times, each time removing a different group and then taking the average for all prediction errors to estimate the test error. The number of groups (v) can be any number from two to the sample size (Hastie *et al.*, 2009).

Bootstrapping involves selecting a random sample with replacement, from the complete data set, until it is the same size as the data set (implying that the sample may contain repeated observations); fitting a model using the bootstrap sample and calculating the predicting error; repeating the sample process a number of times, each time calculating the prediction error; and finally estimating the prediction error by averaging the bootstrap prediction errors (Izenman, 2008).

The expected prediction error calculated with an independent test set is known as the infinite test error or generalisation error. To minimise the generalisation error, the model that fit the data most accurately should not be selected immediately, owing to the possibility that the model is superiorly complicated and in most cases has a small learning error, but a large test error. This leads to the term, 'overfitting' which refers to when a model is too big or complicated, containing too many parameters relative to the learning set's size, resulting in a small learning error as well as a large test (generalisation) error (Izenman, 2008).

3.1.4 Data

Data is referred to as being multivariate when it consists of various observations, measurements or responses on multiple selected variables. When data is easily stored in spreadsheets, it is usually referred to as small, whereas big data sets require special database systems (Izenman, 2008). A database is a collection of data with the primary function to store information. Data consists of many different types of variables, which include:

1. Indexing variables, which generally involve names or serial numbers that identify a variable. Two indexing variables exist, namely a primary key, which uniquely identifies an observation, and a foreign key, which is a primary key in a related database, linked to a variable;
2. Binary variables, which have one of two values, for example '1' or '0', and, yes or no;
3. Boolean variables, which can either be true, false or unknown;
4. Nominal variables, which can be one of a fixed number of string characters, for example toothpaste brands in a shop;

5. Ordinal variables, which are similar to nominal variables by also containing string characters, but which can be ordered, for example excellent, good, neutral and bad;
6. Integer variables, which refer to a non-negative whole number; and
7. Continuous variables, which measured variables containing a certain number of digits and can either be numeric or decimal (Izenman, 2008).

Other variables that exist are fixed, stochastic, input and output variables (Izenman, 2008). The value of a fixed variable has been determined in advance as part of an experiment, whereas a stochastic value is chosen randomly from a list or by another stochastic method. The variables are again grouped in two larger types, namely qualitative (discrete, factors or categorical values) or quantitative (numerical or continuous values).

The difference between an input and output variable is also of great importance. The input variable also known as a predictor or independent variable is generally considered as fixed by means of a statistically designed experiment or stochastic when it refers to uncontrolled observations. The output variable, also known as a response or dependent variable, is stochastic and dependent on the input variable (Izenman, 2008).

3.2 Unsupervised learning

As previously mentioned, unsupervised learning does not have a dependent variable and the methods are mainly descriptive, searching for unexpected patterns or relationships (Bramer, 2007). There are a wide array of methods and only some of the more popular methods will be introduced in this section, because the nature of this project is applicable to supervised learning.

3.2.1 Clustering

Clustering is the most well-known method of unsupervised learning used for analysing multivariate unstructured data. Clustering is also known as data segmentation and class discovery. The algorithm categorises data into two or more groups (clusters) with the main goal of maximising the similarity between the members in the clusters (Jacob & Ramani, 2012).

Clustering and classification methods are similar, but with further investigation, the difference in their philosophies becomes apparent. With classification, the amount of classes or groups present in the data set is known, whereas in clustering, it is unknown. Also, the objective of classification is to classify new items into classes based on rules learnt from the learning set. With clustering, however, no prior information regarding the class structure is available, making it more applicable to data exploration. Lastly, classification almost always deals with classifying observations, whereas clustering can either simultaneously or individually be applied to group observations and/or variables (Izenman, 2008). The majority of clustering methods can be grouped into the following broad categories:

1. **Partition methods** build clusters by dividing the data at certain points, in most cases, ending up with each object belonging to a group.

2. **Hierarchical methods** decompose the data by means of hierarchy, which follow either an agglomerative (bottom-up) or divisive (top-down) approach.

3. **Density-based methods** group objects based on their density. Clusters are formed from data points that occur in the same area and keep growing as long as the density (amount of data points) exceeds some value. This is valuable for identifying outliers.

4. **Grid-based methods** function by restricting the space for the objects to only a number of cells that form a grid structure (Han & Kamber, 2006).

Partitioning methods are the most fundamental clustering method, clustering the data in k-number groups. The most common methods include k-means and k-medoids partitioning. Figure 3.2 gives a representation of data clustered in three groups using k-means. Clustering is evaluated by determining the feasibility of the analysis on the data set; determining the amount of clusters; and measuring the quality. The compactness and separability are also evaluated (Han & Kamber, 2006). Compactness determines how close the elements in the cluster are (small variance) and separability investigates how diverse the cluster is (Veloso *et al.*, 2014).

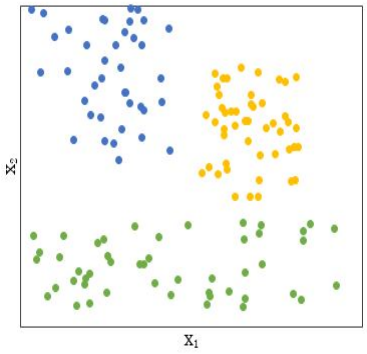


FIGURE 3.2: *A representation of a data set grouped in three clusters using k-means clustering.*

Clustering and association rules have been applied in the healthcare field when similar patient records and related symptoms required investigation (Jacob & Ramani, 2012). A study used clustering techniques to predict readmissions in an intensive care unit. Clusters representing characteristics of readmitted patients were created using three methods, namely k-means, k-medoids and x-means. The three methods' results were evaluated using the Daview-Bouldin index, which found k-means to be the best method, whereas k-medoids got the worst results (Veloso *et al.*, 2014).

3.2.2 Association rules

Association is especially known for being applied to commercial data sets. The aim of the method is to find values that co-exist within A ($A = A_1, A_2, ..., A_n$) and is usually applied to binary data ($A_i \in 0, 1$). For example, A may refer to all the products in the store. When a certain product is bought, the variable for that product, say A_{milk}, is set equal to '1' . This

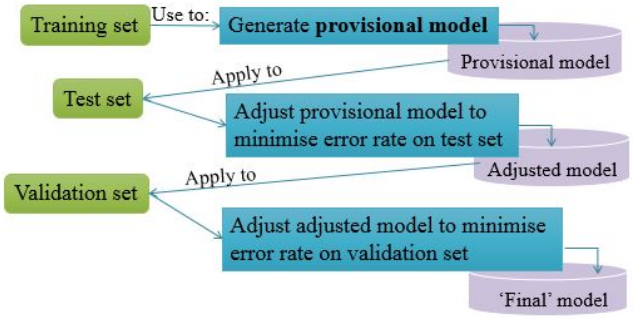


FIGURE 3.3: *General supervised methodology. (Adapted from: Larose (2005).)*

can then be used to determine which products are generally purchased together (Hastie *et al.*, 2009).

Association has ample application in healthcare, for example detecting relationships between diseases, state of health and symptoms. It has also been used for detecting relationships between diseases and drugs as well as fraud and misuse of health insurance. The analysis of healthcare data has been enhanced by integrating classification methods and association rules (Agarwal & Tomar, 2013).

The Apriori algorithm for association rules is commonly used and varies from general association rules by incorporating a minimum contributive requirement and confidence level. In short, the principle is followed that if an item does not exceed the minimum support or is not frequent, it is removed, owing to it not contributing to constructing the association rules. Apriori algorithms have been applied to classify patients with type-2 diabetes; discover frequently occurring diseases; and generate rules for healthy and heart disease patients, i.e. to discover factors causing heart problems in men and women (Agarwal & Tomar, 2013).

3.3 Supervised learning techniques

Similar to unsupervised learning, supervised learning entails a vast number of techniques. The goal of predictive data mining in healthcare is to support clinical decision-making by building models that can use patient data to predict a certain outcome (Bellazzi & Zupan, 2008).

The majority of supervised methods use the algorithm displayed in Figure 3.3 to build and evaluate a model. First, the training set, which comprises pre-classified values for the target and predictor variables, is used to derive a provisional model. Next, this model is evaluated by applying it to the test set, where the values of the independent (target) variables are temporarily hidden. The model is then used to classify the test set, with rules learnt from the training set. The test set's classification accuracy is evaluated by comparing it against the true target variables. The model is then adjusted again to minimise the error rate on the test set and applied to the a validation set, another set that was held out from the start. The model is evaluated and adjusted again to minimise the validation error (Larose, 2005).

Classification is one of the most commonly used data mining methods. A classification example applicable to a hospital may constitute the aim of classifying patients to either have a

high, medium or low risk of contracting a certain infection (Bramer, 2007). There are various classification methods, however the methods are not restricted to only classification problems. For example, support vector machines (SVM), neural networks (NN) and k-nearest neighbour (kNN) methods are also used in regression analysis. The terminology of a problem being either a classification or regression problem generally comes from the nature of the output variable. Predicting qualitative (categorical or discrete) output variables is known as classification, where regression is used with quantitative output variables and in some cases referred to as estimation (Hastie *et al.*, 2009). Generally, in a classification problem, a prediction is made with regard to in which class an observation falls. The methodical process of predicting class is known as a classifier or classification rule (Breiman *et al.*, 1993). Regression is originally a statistical method used for investigating relationship between dependent and independent variables and is further explored in Section 3.6.

The following few subsections are dedicated to some of the more common supervised learning methods. These methods were selected after reviewing literature of data mining methods described in articles related to healthcare as well as cross-validating it with an article on a data mining website about the most popular methods.

3.3.1 Support vector machines

SVMs create a hyperplane that classifies data into two or more classes. SVM is explained in this section for binary classification, for which it was initially developed, but can be used for multi-class problems with multiple hyperplanes for classification as well as regression tasks (Agarwal & Tomar, 2013). Patients have been classified by SVM as having either a low risk or high risk for suffering of diabetes – Figure 3.4 has been constructed to display the basic working of the algorithm on the basis of this example (Agarwal & Tomar, 2013). SVM projects data points onto a higher dimension whilst determining the best hyperplane to separate the data. Kernel functions, for example Gaussion or polynomial, are used for the non-linear mapping of the training sample to the higher dimensional space. The SVM maximises the distance between the hyperplane and the closest data points from the various classes (two classes in the case of a binomial set). These data points which are an equal distance on both sides of the hyperplane are referred to as support vectors. The distance between the closest points on respective sides of the hyperplane is known as the margin. The basic concept of the SVM is to maximise this margin, so that the hyperplane is positioned an equal distance from the respective support vectors (Li, 2015).

SVM is a popular data mining technique, applied to various healthcare problems. Previous studies have used SVM and modified SVM algorithms to classify and identify heart disease and breast cancer (Agarwal & Tomar, 2013). Along with hierarchical clustering, it is commonly applied in analysing voluminous data generated by DNA microarrays and mass spectrometry[1], providing decision support in the field of genomic medicine. In another diabetes study, the researchers used the algorithm to identify the effectiveness of various treatments with different age groups (Aljumah *et al.*, 2012).

[1]Mass spectrometry is used to measure the mass and relative concentrations of atoms and molecules.

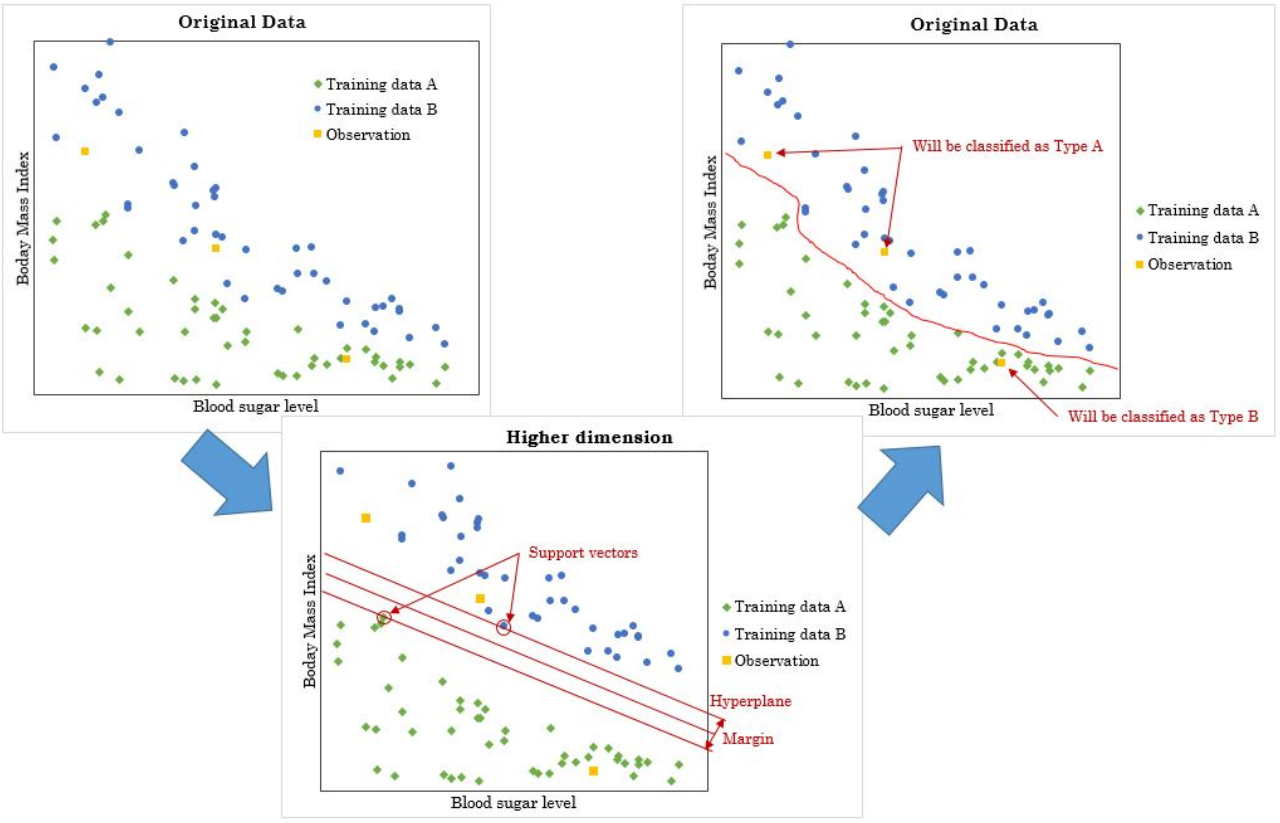


FIGURE 3.4: *The SVM algorithm based on classifying diabetes patients. (Adapted from: Agarwal & Tomar (2013).)*

3.3.2 Neural networks

Neural networks were developed in both statistics and artificial intelligence and are also known as artificial neural networks (ANN). They consist of a two-stage classification or regression model, generally displayed by a network diagram. The main objective is to generate linear combinations from the input data as derived characteristics and then modelling the target variables as a non-linear function of these characteristics (Hastie *et al.*, 2009). Originally, the algorithm was designed as a very simplified model of neuron activity in the brain. Later on, ANN were developed more abstractly, as a network comprising interconnected non-linear connected elements. It is predominantly used for pattern classification and prediction, for example with problems entailing the recognition of speech, faces and characters as well as robotics. These problems entail large sample sizes and high-dimensional data (Izenman, 2008).

In general, the model comprises an input, output and hidden layer, each in turn consisting of nodes (neurons) and links. The input nodes are the predictor variables, whereas the output nodes are analysed as outcome variables. There are numerous additions to the algorithm as with most of the data mining techniques of which the Multilayer perceptron network (MLPN) with backpropagation is well known and powerful in analysing prediction and classification problems (Meng *et al.*, 2012). The input and output nodes are connected and assigned a weight, which is adjusted during the learning phase, so that the network can predict the correct class label of

the input nodes.

NN have been criticised for having high computational costs and long training times; difficulty with interpreting the symbolic meanings of the weights and hidden layers units; and selecting a network structure. On the other side, NN sift out noisy data efficiently and have the ability to classify untrained patterns. It is useful to a researcher who has little knowledge of the relationship between the characteristics and classes and is efficient with analysing continuous input and output variables. (Han & Kamber, 2006).

ANN have been used extensively in clinical medicine which may be as a result of having good predictive performance, modelling complex non-linear relationships better than simpler methods, such as logistic regression, and employing Naive Bayes Models (Bellazzi & Zupan, 2008). In this way, ANN have been implemented to diagnose breast cancer by analysing data from digital mammograms (Jacob & Ramani, 2012). A predictive model has also been constructed with NN to identify high-cost patients in the top five percentile of the general population (Izad Shenas *et al.*, 2014). NN are typically used to recognise handwritten characters, analyse pathology and laboratory data and even develop (train) a computer program to read English text (Han & Kamber, 2006).

3.3.3 Naive Bayes network

Naive Bayes is a combination of classification algorithms that share the assumption that the attributes of the data being classified in a certain class is independent of the other attributes in that class, such as a patient's age and area code. Essentially, the theorem predicts a class based on a set of attributes using probability rules (Li, 2015).

Bayesian networks uses both expert knowledge and data to generate probabilistic graphical models of cause-and-effect relationships between variables as well as supplying a weight to the variables that gives an indication of how probable a variable is to influence another (Paramasivam *et al.*, 2014). Bayesian networks have been used in medical decision-making since the nineties, predicting the probability that a patient has a specific psychiatric disease based on symptoms (Curiac *et al.*, 2009). The development of certain diseases and whether a treatment will be beneficial to a certain patient or not have been predicted.

The performance of Bayesian classification is comparable with decision trees and certain neural network algorithms, exhibiting a high accuracy and calculating speed when used with large databases (Han & Kamber, 2006). An example of this is a 'Bayesian confidence propagation neural network' developed to analyse a database which was the largest of its kind, storing adverse drug reaction data for 47 countries. The aim was to find combinations between drugs, adverse drug reactions and other variables (Bate *et al.*, 1998).

3.3.4 k-Nearest neighbour

k-Nearest Neighbour (kNN) is mostly used for classification problems, but is also used for estimation and prediction. It varies from other data mining methods such as decision trees and SVM by being a 'lazy learner'. The term lazy learner is used when a model essentially only stores the input data during the training phase where other methods will start to build the

classification model (Li, 2015). kNN is an example of instance-based learning and starts by storing the labelled data of the learning set. When unlabelled data is added, kNN classifies the data points by first considering k number closest labelled data points (or neighbours) and uses their information to decide on a class for the new data (Li, 2015).

The distance between the labelled and unlabelled data is determined with a function that accounts for three aspects: the distance is always non-negative; the distance between say point 1 to point 2 is the same as point 2 to point 1 (commutativity); and when a third point is introduced between point 1 and 2, it will not shorten the distance between two other points (triangle inequality) (Larose, 2005). The most commonly used distance function is the Euclidean distance, which is based on how humans measure distance. To account for very large attributes (such as income) very small attributes (such as age) the values are normalised. However, the Euclidean distance is not suitable for categorical variables, for which a function has to be defined instead (Larose, 2005). After a method for measuring distance is established, a classification decision has to be made. This can either be achieved with unweighted or weighted voting. The first is the most simple voting method which entails deciding the value of k (the amount of records that have an influence on the classification decision), comparing the new record with the k nearest neighbours and then classify the record based on one vote for each nearest neighbour. Weighted voting follows the same steps, except that neighbours are weighted inversely proportional to the distance from the new point, resulting in closer neighbours having a greater weighting and minimising the likelihood of ties (Larose, 2005).

kNN is a popular method owing to interpretation and implementation being simpler than other methods and, depending on the distance metric, quite accurate. However, kNN can get computationally expensive with large data sets; is not robust regarding noisy data; and, in the case where some attribute has a large range and other smaller ranges, it must be scaled (Li, 2015).

3.3.5 Decision trees

Decision trees are a well-known data mining method because they can be interpreted easily and possess high computational accuracy. The method entails recursively partitioning the data set into discrete subcategories, based on the value of a certain variable (Paramasivam *et al.*, 2014). A decision tree, for example displayed in Figure 3.5, comprises decision nodes, which are connected by branches, extending from the root node, which is usually at the top of the tree diagram, towards terminating leaf nodes. Variables are tested from the root node at each decision node, with the possible outcome being represented by a branch, which again leads to another decision node or terminating leaf node. When the tree cannot be split further, no new nodes are grown (Larose, 2005). For example, in Figure 3.5, the node for where savings are equal to medium, all the instances are classified to be good credit risks, resulting in a 100% pure node and no further splitting options are available. This is not always the case and therefore there are various methods to measure purity and decide on a cut-off value. Two leading decision-tree algorithms, namely classification and regression trees (CART) and C4.5, will be introduced in this section.

Problems to be modelled by decision trees have to be supervised learning problems, with pre-classified variables that can be used for the learning set. The learning set must also have a wide array of data points representing the data points that will require classification in the future.

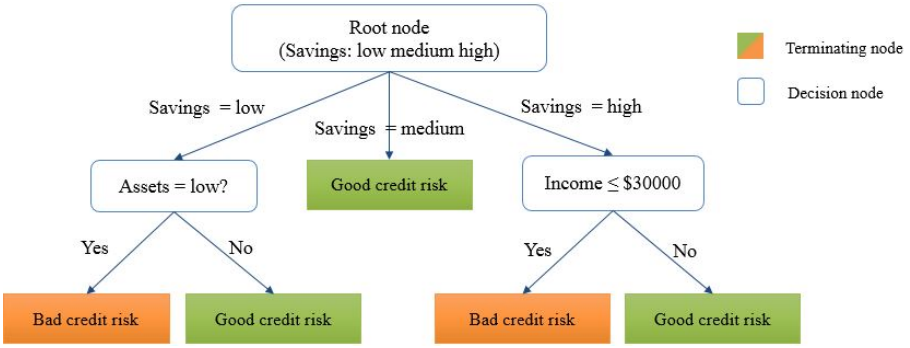


FIGURE 3.5: *An general example of a decision tree to determine whether a potential customer being is a good or bad credit risk. (Adapted from: Larose (2005).)*

Decision trees rely heavily on learning from example (Larose, 2005). Tree-based methods, such as CART and C4.5, have a wide application field, especially in applied science, political science, speech recognition, marketing and biomedical and genetic research (Izenman, 2008). Decision trees have been used to predict the chance of survival for a patient suffering from breast cancer and to characterise types of skin conditions in adults and children (Agarwal & Tomar, 2013). In another case, a hybrid tree was developed to classify the activities of a patient suffering from a chronic disease (Agarwal & Tomar, 2013).

3.3.5.1 Classification and regression trees

CART is a nonparametric statistical method that mainly uses a recursive partitioning algorithm. Recursive partitioning is a stepwise process to the formation of a decision tree by either splitting or not splitting a node into two child nodes. One of the factors contributing to the popularity of the CART method is that the results can be interpreted and understood easily owing to the algorithm asking hierarchical boolean questions in sequence. Classification and regression are both supervised data mining techniques, but differ in their output variables. For binary classification models, the dependent variable (Y) has a binary value, whereas with regression problems, Y is continuous (Bramer, 2007).

This method was first developed by (Breiman *et al.*, 1993) who described it as binary tree-structured classifiers. The name 'CART' came from the computer program that was used to solve the algorithm. An example of a six-class tree can be seen in Figure 3.6. From the figure, it can be seen that $G = G_2 \cup G_3$ and in the same way $G_3 = G_6 \cup G_7$. Subsets that are not split are referred to as terminal subsets (rectangular nodes). A class label (j) is given to each terminal node, where two or more terminal nodes may have the same allocated class label. The classifiers are determined by adding terminal nodes that belong to a certain class, e.g. $j_1 = G_{15}$ and $j_4 = G_6 \cup G_{17}$. The splits are determined by conditions of the measurement vector ($\boldsymbol{g}$) for example, the split into G_2 and G_3 may be in the form of $G_2 = \{\boldsymbol{g};\ g_4 \leq 7\}$, $G_3 = \{\boldsymbol{g};\ g_4 > 7\}$ (Breiman *et al.*, 1993).

The tree method predicts a class for $\boldsymbol{g}$ by determining in which subset $\boldsymbol{g}$ goes, for example at split 1, $\boldsymbol{g}$ goes into G_2 if g_4 is smaller or equal to seven and so forth. When $\boldsymbol{g}$ reaches a terminal subset, its predicted class is the class label of that subset.

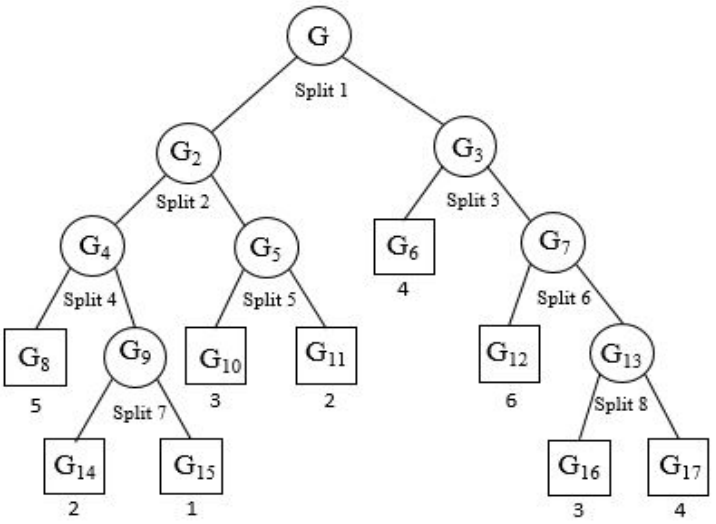


FIGURE 3.6: *Example of the classification tree developed for a heart attack study. (Adapted from: Breiman* et al. *(1993).)*

Constructing a tree comprises mainly of three steps, namely:

1. Selecting the splits;
2. Deciding when to stop splitting and declare a terminal node; and
3. Assigning a terminal node to a class.

Some statistical techniques are designed for homogeneous, small and structured data sets where the variables are of the same type. Data is homogeneous when the relationship between the variables are the same over the measurement space. CART, in turn, is designed for larger data sets, with higher complexity, where complexity is characterised by high dimensionality, different data types, having a non-standard data set and, of most concern, non-homogeneity.High dimensionality refers to the data points being sparser and spread further apart. CART is described in more detail in Section 3.7.

3.3.5.2 C4.5

The C4.5 decision-tree algorithm is the successor of the ID3 algorithm, which was first developed for generating decision trees, and the predecessor of the C5 method, although the C4.5 is more applied. Similar to CART, the C4.5 also uses recursive partitioning; however, there exist fundamental differences between them when analysed in more detail (Larose, 2005). Firstly, C4.5 is not restricted to binary splits, such as CART, and for categorical variables, it produces a separate branch by default for each value of the categorical variable. This results in a 'bushier' tree which is not always ideal, owing to the some groups having a low frequency (Larose, 2005). The C4.5 algorithm used 'information gain' (entropy) to determine the splits, whereas CART uses the more simple Gini impurity index.

Overfitting or pruning is accounted for by a single pruning process with the C4.5 method, whereas CART uses a cost-complexity method. The cost complexity method starts at the bottom and evaluates the misclassification cost of the tree with and without a certain node (Li, 2015). Error rates are also calculated differently, with CART employing cross-validation and C4.5 using a heuristic formula. Missing values are handled using probabilistic distribution with C4.5, whereas CART incorporates surrogates (Li, 2015).

3.4 Data mining in healthcare

Data mining methods are not new in the field of healthcare and are not only becoming more popular, but may also be becoming more essential as the healthcare sector experiences increasing pressure to reduce costs while also improving the quality of care.

Utilising healthcare information systems is one of the strategies to address the issues related to healthcare facilities in general, collecting and generating vast amounts of data. Although analysing the available data for a specific problem can be vast and complex, it may lead to improved decision-making (Krishnan, 2010).

Medical data is collected from humans and accordingly there exists a large legal and ethical tradition designed to prevent patients from being abused by misusing the data. The main legal, social and ethical aspects in medicine can be organised into five groups, namely administrative issues, privacy and security of human data, expected benefits, fear of lawsuits and data ownership (Krishnan, 2010).

Clinical data mining specifically focusses on discovering new information and patterns which may be used to improve administrative and medical decision-making, evaluate insurance policies, support clinicians in diagnoses, prognoses and treatment methods, and predicting diseases (Agarwal & Tomar, 2013). Other applications of data mining in the healthcare sector include:

- Constructing models for managing hospital resources, e.g. to detect chronic diseases and prioritise patients in order to treat patients timely and effectively;
- Ranking hospitals on their ability to treat high-risk patients or handle emergency situations;
- Improving customer relationships by identifying specific customer needs, patterns, preferences, and appeal to potential customers;
- Identifying irregular patterns in infection control data and control future or wide spread infection in the hospital;
- Differentiating between treatment techniques by analysing the effectiveness of available treatments, identifying side effects, reducing risks and developing better treatment methodologies;
- Analysing patient data, building predictive models and identifying current and future patient requirements and preferences can improve patient care and satisfaction;

- Developing models that detect fraud and abuse of medical claims can decrease insurance fraud in healthcare;

- Identifying high-risk patients, for example ones suffering from diabetes; and

- Planning health policies effectively: making them more cost-effective and improving health quality (Koh & Tan, 2005).

Examples of data mining methods used in healthcare have been mentioned whilst introducing the more popular data mining methods. However there exist ample more examples of data mining models applied to problems in the healthcare sector, with many algorithms incorporating a combination of models for improved results.

A study have been conducted to compare twenty classification algorithms' performance on analysing a data set comprising 155 instances and 19 attribute variables about the cancerous Hepatitis C virus. The random tree classification and the C4.5 algorithm had the highest classification accuracy (Jacob & Ramani, 2012). Another study, which also compared various methods by classifying DNA sequence records in one of three classes, found that random forest classification and C4.5 methods were the best. The same data scientist also compared methods on a diverse clinical set comprising of patients suffering from various diseases and health problems. Binary logistic regression, cost-sensitive classification, C4.5 and random forests performed the best (Jacob & Ramani, 2012). A study that extracted rules for prostate cancer patients used a clustering method to group them in either low-, medium- and high-risk classes. A detailed study on clustering methods for clinical data sets found that between hierarchical-, partitioning-, density- and grid-based clustering algorithms, k-means and hierarchical agglomerative clustering are more suitable for clinical databases. HIV-infected patients' data records have been mined using an n-cross validation Apriori algorithm (Jacob & Ramani, 2012).

One of the most challenging aspects of data mining in the healthcare field is obtaining relevant, complete and quality data. The data is heterogeneous and complex as it is compiled from different sources such as laboratories, doctors and administration (Agarwal & Tomar, 2013). Research has suggested building a data warehouse before mining the data, but it is expensive and time-consuming. Without adequate knowledge of the field, the pitfalls of data mining might not be clearly understood and thus a data mining team should have knowledge about the domain as well as experience (Koh & Tan, 2005). Healthcare institutions that are developing data mining applications are required to invest a lot of money, time and effort. In general, projects fail due to a lack of support from management, unrealistic user requirements, lack of data mining expertise and poor project management. Healthcare practitioners and managers have to participate in the project and also be convinced of the value of data mining and willing to change work processes if required (Koh & Tan, 2005). Data mining entails the application of many varying techniques from interdisciplinary fields and is often seen as difficult to learn and more so to master. There may also be more than one technique that serve the same purpose and behave equally good, but not feasible to apply all the alternatives, leaving the method choice to the instinct of an expert (Bellazzi & Zupan, 2008).

3.5 Similar published studies investigating readmission

The studies, which were introduced in Chapter 2 by describing the objectives, sample size and study period, were also analysed to determine which statistical or data mining methods were incorporated. The studies, described in Table 3.1, predominantly used Cox regression analysis and logistic regression to determine the indicators for readmission. Chi-square analysis was used to test the variables that may result in readmission, as well as compare two groups for example, patient that are readmitted and not readmitted. Variables were regarded as significant at p-values of less than 0.05 in most of the studies.

TABLE 3.1: *Methods used in various published studies to investigate readmission.*

Citation	**Method**	**Chi-square** [Chi]	**Logistic regression** [Log]	**Survival analysis** [Surv]	**Multiple regression** [Multi]	**Other**
(Barekatain *et al.*, 2013)	**Negative binomial regression** determines factors associated with the number of readmissions. The dependent variable (number of readmissions) is numeric, not normally distributed and has a small variance. Significance level: $p<0.05$.					1
(Bernardo & Forchuk, 2001)	**Chi-square tests** (nominal data) and **t-tests** (continuous data) examine the differences between readmitted patients and those not readmitted. **Correlation analysis (Pearson's coefficient)** and **multiple regression analysis** predict the expected number of readmissions as well as the factors responsible for readmission. Significance level: $p<0.05$.	1			1	1
(Byrne *et al.*, 2010)	**Multiple regression analysis** investigates the predictors for readmission within 30 days, six months and five years and further compares the predictors of readmission for patients with a number of diagnoses, multiple admissions and first-time admitted depressed patients.				1	1
(Durbin *et al.*, 2007)	**Systematic review** of bibliographic indexes to determine if indicators of early readmission are associated with the quality of inpatient psychiatric care.					1
(Gillis *et al.*, 1986)	**Professional findings from descriptive statistics**: Healthcare specialists investigate factors that affect readmission. **Logistic regression** investigate readmission, but eliminates possible interaction between variables.		1			
(Haywood *et al.*, 1995)	**Logarithmic transformation** normalises the gathered data. **Chi-square** categorically analyses which variables are generally associated with the amount of admissions. **Mantel-Haenszel test** determines whether the amount of admissions increases (or decreases) as a function of each predictor variable. Lastly, **multiple regression** was applied to examine the predictive strength of the variables.	1			1	1
(Heggestad, 2001)	**Cox regression** investigates the relationship between hospital practices and likelihood for readmission. **Log-log survival curves**: assess the proportional hazard assumption. **Stepwise regression** conducted as a secondary test. Significance level: $p<0.05$.			1		1

Methods used to calculate readmission rates continued.

Citation	Method	[Chi]	[Log]	[Surv]	[Multi]	[Other]
(Heslin *et al.*, 2015)	Literature review of studies conducted in 2012 in the United States of America.					1
(Innes *et al.*, 2015)	**Cox regression** analyses the relationship between variables and resulting risk of a depression-related admission (and, secondly, risk of readmission) with **hazard ratios** for age, gender and residential deprivation quintile. Significance level: $p<0.05$.			1		
(Johnstone & Zolese, 1999)	**Identify and review studies** of randomised controlled trials about planned short stay versus long stay for patients with serious mental illnesses. The data was analysed and compared using **Peto odds ratios**. Significance level: $p<0.05$.					1
(Jones *et al.*, 2002)	**Cox regression** investigates six diagnostic categories as predictors for readmission (Software: SPSS 10.0). **Survival plots** are used for visual comparisons of the survival functions for each diagnosis group. **Forward stepwise Cox regression** determines whether sociodemographic factors predict readmission and if they interact with the type of diagnosis.			1		
(Loch, 2012)	**T-test (continuous variables) and chi-square tests (categorical variables)** compare the variables between two patient groups. **Backwards stepwise logistic regression** investigate the relationship between the indicators and outcome variables. **Cox proportional hazard regression model** and **logistic regression** determine the variables that predict time to one-year readmission the best. **Kaplan-Meyer survival curves** generate two graphs with 'rehospitalisation survival' as outcome variable. **Log-rank** estimates the statistical difference between the two graphs.	1	1			1
(Lyons *et al.*, 1997)	**T-tests and logistic regression** analyse variables and predict readmission within one year. **Hierarchical multiple regression** determines whether patients discharged prematurely were more likely to be readmitted.		1		1	
(Malesu)	The methods are not presented in the text.					
(Mark *et al.*, 2006)	The number of patients with detoxification readmissions within 30, 60, 90, 180, and 365 days of initial admission was determined and the factors were compared to patients not readmitted. **Cox proportional hazard model** determines the influence several factors have on the time to detoxification readmission. (Software: SAS PHREG).			1		
(Mayoral *et al.*, 2012)	**Chi-square test** detects relationships between different factors and readmission (yes or no). **Logistic regression** builds a multiple variable model to quantify factors that may result in readmission. Significance level: $p<0.05$.	1	1			
(Moss *et al.*, 2014)	**Cox regression** determines and analyses variables associated with a shorter time to readmission. **Backward step-wise Cox regression and hazard plots** determine most significant variables and model the factors. Cox regression was also used to determine relevant covariates (Software: IBM SPSS Statistics 20.0).			1		
(Niehaus *et al.*, 2008)	**Kaplan-Meier survival curves** graphically depict the influence of crisis-discharge on the time to readmission. **Cox proportional hazards regression** tests if there is a difference in the time to readmission between non-crisis-discharged and crisis-discharged patients. (Software: R)			1		

Methods used to calculate readmission rates continued.

Citation	Method	[Chi]	[Log]	[Surv]	[Multi]	[Other]
(Wickizer & Lessler, 1998)	**Logistic regression** investigates the effect of restrictions on the length of stay in comparison to 60-day readmission rates. The sample included a limited set of variables, which were used as covariates in the logistic regression analysis. **Chi-square test and bivariate analyses** test the difference or similarities between the requested and approved days for the various types of diagnosis.	1	1			
(Yussuf *et al.*, 2008)	**Chi-square, frequency distribution and cross-tabulation** analyse categorical variables. **Analysis of variance (ANOVA)** compares means of the continuous variables. **Pearson's correlation** correlates the variables with significant association to readmission. **Logistic regression** determines the predictive strength of the significant variables. Significance level: $p<0.05$. (Software: SPSS version 11).	1	1			1
		6	6	6	4	10

From the studies summarised in Table 3.1, methods were identified that can be incorporated to analyse the data set pertaining to this project. It will also serve as comparison with regard to other methods identified in literature and proposed by statistical experts.

3.6 Regression analysis

Regression describes the relationship between a dependent variable and one or more independent variables. This description, however, assumes there is a relationship, with the fit being either good or poor. The model can involve simple regression (straight line), curvilinear regression (curved line), multiple regression (dependent variable predicted by two or more independent variables) and multi-variable regression (more than one dependent variable) (Riffenburgh, 2012).

The relationship between X and Y can take on various forms ranging from straight lines to curved models such as a parabola, third-degree, logarithmic, exponential, biological growth curve and sine waves (Riffenburgh, 2012). Theoretical or population coefficients of the models are denoted by the symbol $\mathcal{B}$ and sample estimates are denoted by b. For a straight line model, including simple regression, the model is determined by two pieces of information. The best fit is calculated using least squares

$$y - \bar{y} = b_1(x - \bar{x}) = \frac{s_{xy}}{(s_x)^2}(x - \bar{x}). \tag{3.1}$$

The slope of the regression line, b_1, is calculated by dividing the covariance of x and y (s_{xy}) by the variance of x. If multiple values of y are to be predicted from x, (3.1) can be modified using the slope-intercept form as

$$y|x = b_0 + b_1 x, \qquad \text{where} \quad b_0 = \bar{y} - b_1\bar{x} \quad \text{(Riffenburgh, 2012)}. \tag{3.2}$$

Background pertaining to fitting a straight line model from data points using either the intercept and slope, or mean and slope is briefly discussed in Appendix B.2.1. There are five general

assumptions when attempting regression:

1. The errors in the data points, the deviations from the average, are independent from each other;
2. The the model used to fit the data set is appropriate;
3. The independent data points (x) are measured without error as exact known values;
4. The variance in the dependent variables (y) is the same for all values of the independent variables (x); and
5. The distribution of y is more or less normal for the values of x (Riffenburgh, 2012).

Regression is generally robust, especially if assumptions four and five are violated, and, unless the violations are unreasonable, they should not pose a problem (Riffenburgh, 2012). Assumption three is important owing to the least-squares method used in regression minimising only the sum of squared errors vertically (y-axis). When the error on the x-axis is too large to ignore, Deming's regression can be used to account for this (Riffenburgh, 2012). The 'correlation error', as it is referred to in this case, depends on knowing the ratio (λ) of the theoretical variances in the individual measurements of x and y,

$$\lambda = \frac{\text{var(a y reading)}}{\text{var(a x reading)}}, \tag{3.3}$$

from which it is clear that λ is difficult to determine. If the error in x is small, simple regression is usually used and if the error is thought to be large and similar to the error in y, λ is taken as one (Riffenburgh, 2012). The ratio can also be estimated after which the regression coefficient (b_1) is calculated by

$$b_1 = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2s_{xy}}. \tag{3.4}$$

The symbols s_y and s_x are the standard deviation in the y and x observations respectively. The covariance of x and y is denoted by s_{xy}.

3.6.1 Investigating the regression model

Generally there are two tools for assessing regression, namely the coefficient of determination (R^2) and the standard error. R^2 is the proportion of points possibly perfectly predicted by the model. With simple regression, it is calculated by squaring the correlation coefficient, whereas with curvilinear and multiple regression, it is a bit more complicated. R^2 is also interpreted as the percentage of x's that describe y correctly (Riffenburgh, 2012).

Standard error (s_m) is the standard deviation of a statistic and is calculated for (i) the residuals (s_e) (the deviation between the observation and the regression line), ; (ii) the estimate of the

regression slope b_1 (s_b); (iii) the estimate of the mean values of y for each x ($s_{\bar{y}|x}$); and (iv) the individual predictions of y for each x ($s_{y|x}$). The sum of squares of the deviations from the regression line is calculated as

$$s_e = \sqrt{\frac{n-1}{n-2}\left((s_y)^2 - (b_1)^2(s_x)^2\right)} \quad = \quad s_y\sqrt{\frac{n-1}{n-2}(1-R^2)} \tag{3.5}$$

The sample size is denoted by n and (3.5) is calculated at $n-2$ degrees of freedom (Riffenburgh, 2012).

Five questions are usually asked about the regression model (Riffenburgh, 2012):

1. Is y significantly predicted by x?
2. How strongly does x predict y?
3. What are the confidence limits on y for a certain value of x?
4. What are the confidence limits on the best predicted y for a certain x?
5. If two samples are estimated, do the regression slopes vary?

If the slope of the regression line is horizontal, the predicted y is the same for all x values, indicating no prediction ability. To test this, a hypothesis test can be conducted by comparing a calculated 'test statistic' (t_{TS}) to a critical t_{crit} value from the t-distribution:

$$H_0 : \mathcal{B} = 0, \text{ opposed to}$$
$$H_1 : \mathcal{B} \neq 0 (or > 0 \text{ or } 0 >)$$

with the standard error of the slope

$$s_b = \frac{s_e}{\sqrt{n-1}s_x}, \tag{3.6}$$

and t_{TS} at $n-2$ degrees of freedom

$$t_{TS} = \frac{b_1}{s_b}. \tag{3.7}$$

If t_{TS} is within the bounds of the critical value, there is no evidence to reject H_0, which indicates that the slope is not a straight line and that x predicts y significantly. The slope can also be compared to a theoretical slope ($\mathcal{B}_1$), which may be the case if a certain relationship is expected or comparing to previous published results (Riffenburgh, 2012).

For the second question (how strongly does x predict y), the predictive ability of the model is indicated by the coefficient of determination R^2, which is generally used in simple regression (straight line model). The closer R^2 is to zero, the more evidence there is of no relationship,

where one is perfect predictability (Riffenburgh, 2012). In some cases, a confidence interval (CI) on the slope of the regression line is informative, with a tight CI suggesting a strong relationship. Information on calculating the confidence intervals for various elements of the regression model and comparing the slopes of two estimated samples is presented in Appendix B.2.2.

Correlation is a popular term in statistics where the correlation coefficient (r) is indicative of the level of association and measured by how closely or loosely the observations (x, y) group around the regression line. The correlation coefficient (r) is defined between -1 and 1, where the sign indicates either a positive or negative relationship with -1 or 1 indicating a very strong (positive or negative) relationship and r closer to zero ($|r| = 0.2$) indicating a weak to no (r=0) relationship (Riffenburgh, 2012). Correlation analysis is further introduced in Appendix B.2.3.

3.6.2 Types of regression

There are various types of regression models available for the different classes of outcomes, as shown in Table 3.2 (Riffenburgh, 2012). Take note that the classes only refer to the variable being predicted, and not the predictors.

TABLE 3.2: *Types of regression models. (Adapted from: (Riffenburgh, 2012).)*

Nature of dependent variable	**Regression model**
Continuous	Ordinary
Rank	Ordered
Categorical: two	Logistic
Categorical: several	Multinomial
Count (# occurrences)	Poisson
Survival (time)	Cox (proportional hazard)

Riffenburgh (2012) identifies the following types of regression and their uses:

- Ordinary regression might be used to predict antigen levels from the body mass index;
- Ordered regression might be used with insomniacs being asked to rank sleeping aids;
- Logistic regression can be used to predict whether breast cancer recurred after post-radiation therapy (yes or no);
- Multinomial regression might be used to classify patients after a year of treatment as either cured, recurred, progressed and so forth;
- Poisson regression can be used to predict the amount of infectious diseases over a two-year period by considering supplements, antibiotics and exercise levels; and
- Cox regression can be used to determine the survival, depending on time, predicted by certain treatments with the outcome being either survival or death.

3.6.3 Multiple and curvilinear regression

Regression analysis of a dependent variable and several independent variables is known as multiple regression. Similar to fitting a model with a line, a curved shape might fit the data better and

may take any form, for example a parabola, logarithmic shape or cyclic pattern (Riffenburgh, 2012). Multiple regression is similar to simple regression, but as it deals with more independent variables, the plane is extended to x_1, x_2, y or in the case of a hyper plane, to $x_1, x_2, x_3, ...y$ (Riffenburgh, 2012). The model is also not confined to a plane, but can also be a curved hyper-surface, where the variables are not all first degree (for example, $y = \mathcal{B}_0 + \mathcal{B}_1 x_1 + \mathcal{B}_2 x_2 + \mathcal{B}_3 x_2^2$).

3.6.3.1 Multiple regression

There is not one method to develop a multiple regression model. A physiological model that suggests which variables to include will be ideal, but in some cases, the model has to be developed by identifying variables to be included as well as their individual contribution, using various regression techniques. For example, univariate regression can be used to test a potential variable and add it to the model if it will make a contribution. Another method may be to include all potential variables and remove those that do not make a contribution. Both these options are a form of stepwise multiple regression (Riffenburgh, 2012).

The regression curve is the result of some or other 'best' criteria. With simple regression, least squares is used to determine the criteria. Other criteria may be either a combination of unbiasedness, minimum variance or maximum likelihood, to name a few. With the more complicated models, the linear equations are solved simultaneously with matrix algebra, best calculated with mathematical computer programs that form part of statistical packages. The best fit is calculated after which an F-test is conducted to determine the significance of the model and the resulting p-value (Riffenburgh, 2012). The software packages calculate various tests and information, which are most likely used for:

1. Validating the model overall and identifying relationships, for example whether the relationship between the dependent variable and the predictors are real or due to fluctuations in the sample;

2. Determining the predictive capability of the model from R^2;

3. Evaluating the contribution of the variables and ranking them according to their predictive capability from the p-values of the t-tests;

4. Identifying the clinical usefulness of predictors – if a component is suspected not to be predictive (maybe has the smallest test statistic or largest p-value), it is removed and the regression recalculated. The reduction in R^2 with the removed component indicates how useful the variable is as a predictor. If the p-value is small and the R^2 difference large enough, the component may also be removed; and

5. Developing a model and prediction equation (Riffenburgh, 2012).

As briefly mentioned before, in the case where predictor variables are added one by one, a form of forward stepwise regression is followed. The coefficient of determination (R^2) can be monitored to evaluate the predictive capability of each variable added. This is only possible if the variables are added in the order of their individual predictive capability which is determined by single

predictor analyses and that the addition of variables are stopped as soon as an unimportant level of predictive ability is added by a variable (Riffenburgh, 2012).

Backward stepwise regression is more commonly used. It entails including all the variables and eliminating the least contributing variables one by one until the predictive capability of the model reduces excessively (Riffenburgh, 2012). Statistical software can conduct backward stepwise regression, but control is lost over overlapping correlating variables that are removed. The software may correct the correlation error between two variables by removing a variable only because another variable co-existed at that time (Riffenburgh, 2012). One benefit of the process is that the model cannot be manipulated by the researcher. With the software, the cut-off value for retaining or removing a variable must be specified, where a variable with a p-value larger than cut-off value is removed (Riffenburgh, 2012).

Nominal data values, which cannot be ordered (for example disease types), are the only variables that require manipulation for use in multiple regression. With two classes, such as male or female, no alteration is required, but with more than two classes, dummy variables should be used (Riffenburgh, 2012). A variable for each possible class should be created, with 1 indicating a data point as part of the class and 2 as not. As a result, for j-classes, $j-1$ variables are used to replace the original variable.

3.6.3.2 Curvilinear regression

Curvilinear regression is essentially the same as simple regression with the difference that the model does not have to be a straight line (Riffenburgh, 2012). The mathematics are similar to that used for multiple regression by also starting with a univariate straight line regression, but instead of adding a second variable, the first variable is squared and used in the second position. The aim of the study determines the model used, which can be derived from theory; comparing it to an already established form; or, in the case of prediction, the model might be suggested by the shape of the data plot (Riffenburgh, 2012). The analysis, interpretation, stepwise methods and use of nominal variables are the same as with multiple regression.

3.6.4 Survival analysis and logistic regression

The term survival is used more broadly as merely not dying. To name two examples, it may refer to the success of a treatment or if a patient was rehospitalised or not. Survival analysis may be conducted to analyse a binary variable, for example, whether a patient survives or not, and which variables are associated to surviving. An example of such an analysis method is logistic regression, which is discussed in Section 3.6.4.4. Survival may also investigated as a time-dependent variable: for what length of time has the subject survived.

3.6.4.1 Estimating survival times

Survival time data are times to an event (death). The proportion of survivors in the group is calculated at successive time points, with the information used to form a 'life table'. An example of a life table representing men older than 45 years diagnosed with diabetes is displayed in Table 3.3. With most data sets some members are lost to follow-up before death occurred, thus no

survival times are recorded (Riffenburgh, 2012). The information on these members does not have to be deleted completely and can be included in the set while 'alive' and then removed from the database when they are lost, whilst adjusting the baseline total number of participants in order to calculate the correct proportion. The data is then referred to as censored. The term censoring refers to a value which is not fully known. Data is right censored when the patient leaves the study before an event occurred, or the study ends before an event occurred. Left censoring occurs when the event took place before the study started.

TABLE 3.3: *Life table of men suffering from diabetes. (Adapted from: Riffenburgh (2012).)*

Interval (years)	Begin	Died	Lost	End	Survived
0	319	0	0	319	1
>0-2	319	16	0	303	0.9498
>2-4	303	19	0	284	0.8902
>4-6	284	19	2	263	0.8306

Survival time data is commonly displayed with survival curves and specifically, Kaplan-Meier graphs. These graphs are preferred to the life table because they are more accurate and easily generated with statistical software (Riffenburgh, 2012). The graphs also keep track of the lost cases. Statistical software is further used to determine the confidence interval on survival estimates for each of the time intervals, leading to a confidence curve of similar shape to the Kaplan-Meier graph.

In some cases, two survival curves might be compared to determine whether the difference is significant or not. An example is to compare survival graphs of females with diabetes to men suffering from diabetes. One commonly used method is the log-rank test which implements the chi-square statistic based on the difference between the expected survival, calculated as if the two graphs were the same, and the observed survival (Riffenburgh, 2012). This is similar to the chi-square goodness-of-fit test, but the log-rank uses matrix algebra and is therefore more complicated. Alternative tests which can also be considered are Cox proportional hazard test and Mantel-Haenszel test. The latter is similar to the log-rank test, but is restricted to comparing only two graphs. Cox regression allows the risk of death to change within the model, whereas the risk is assumed constant with the log-rank test. When using Cox regression it is advised to get help from a statistician as the mathematics can get complicated (Riffenburgh, 2012).

3.6.4.2 Predicting survival time using Cox regression

Survival time analysis addresses the time-to-event aspect focussing on how long an data entry, for example a patient, survived. Cox proportional regression evaluates the proportion between the 'time while at risk' versus the 'level of risk' (hazard) accumulated over time. The probability of a 'death' occurring, assuming it has not yet happened up to a certain time (t), is referred to as a hazard and denoted as $H(t)$ (Riffenburgh, 2012). The hazard that will take place if none of the potential independent variables were present is known as the baseline hazard ($H_0(t)$) and the hazard ratio is the ratio between the hazard with the occurrence of an independent variable and the baseline hazard $H(t)/H_0(t)$. Proportional hazards assume that predictors affect a hazard proportionally, for example increasing the risk by 15%.

This model makes use of two groups, for example a treatment and control group, and assumes that the hazards for the groups are proportional regardless of the baseline or length of time after the baseline. The hazard ratio is thus independent of time and the influencing factor's effect can be calculated without the probability distribution of individual hazards being known. The Cox model comprises

$$H(t) = H_0(t) \exp^{b_1 m_1 + ... + b_k m_k}, \tag{3.8}$$

dividing (3.8) by the baseline hazard and computing the log results

$$\ln[H(t)/H_0(t)] = b_1 m_1 + ... + b_k m_k. \tag{3.9}$$

Similar to the method used with logistic regression, the right side of (3.9) is the same as ordinal multiple regression and the regression of the log hazard ratio can be calculated for the set of independent variables (Riffenburgh, 2012). Cox regression calculates the estimates of hazard ratios, reflecting survival experience, where logistic regression's odds ratio is related to event experience (Lemke, 2012). Software can compute Cox regression and after computing the b's, the probability for survival at any given time ($H(t)$) is generally given as

$$H(t) = \exp^{-[H_0(t) \times (b_1 m_1 + ... + b_k m_k)]}. \tag{3.10}$$

Cox proportional hazards assume continuous readings on the independent variables in order to prevent ties for occurring. There are two forms by Breslow and Efron respectively that provide adjustment methods in the case where data is discrete. The software outputs that are of importance are the overall p-value for the model, each predictor's p-value and confidence level as well as the hazard ratio (Riffenburgh, 2012).

3.6.4.3 PROC PHREG method for estimating 30-day readmission

The SAS/STAT PHREG procedure, employing Cox regression analysis, was used to model time to readmission whilst also adjusting for covariates, including time censoring and a decrease in patients. The impact of recurrent readmissions and time-varying covariates were also analysed (Lemke, 2012). The study and methodology are introduced in Appendix B.2.4. The study found that Cox proportional hazard regression gives valuable insight with regard to modelling the risk for 30-day readmission by including fixed-value and time-varying predictors (Lemke, 2012).

3.6.4.4 Logistic regression

Logistic regression is implemented if the outcome variable is binary, for example survival (1) versus failure to survive (0). For a binary y, the predicted value is the probability that a survival (1) will occur and is estimated by the proportion of 1's in the sample. The sample proportion (p_m), which is the sum of 0's and 1's divided by the sample size, is used to calculate the chance for survival. The logarithm of the aforementioned 'chance' is known as the log odds ratio and used to calculate statistical tests when the distribution of the log odds ratio is known (Riffenburgh,

2012). Generally regression methods are followed after the dependent variable is transformed to the log odds ratio by

$$\ln\left(\frac{p_m}{1-p_m}\right) = \mathcal{B}_0 + \mathcal{B}_1 x. \tag{3.11}$$

In essence, multiple and curvilinear logistic regression are implemented similarly to ordinary regression, with only the dependent variable transformed. The right side of (3.11) can be transformed to include other terms as required with simple, multiple or curvilinear regression (Riffenburgh, 2012).

In the case where a model is validated or explored to identify relationships, the p-value will provide an indication of whether the relationship between y and the model is real or due to sampling fluctuations. With logistic regression, a chi-square test, instead of an F-test, is conducted (Riffenburgh, 2012). Most statistical packages provide an estimate for the coefficient of determination (R^2) that will help indicate whether the predictive capability of the model is useful. However, because of the transformation, software usually provide an estimate for R^2. Normal (z) tests can be used to evaluate each component's contribution and rank them according to clinical significance by calculating the p-values. The individual component's predictive capability can also be evaluated by monitoring the change in the R^2 equivalent or chi-square value when changing the model (Riffenburgh, 2012).

3.7 CART

As briefly mentioned in section 3.3.5.1, CART is a nonparametric statistical method that systematically forms a decision tree by either splitting or not splitting a node into two child nodes. With classification models, the dependent variable (Y) has a binary value, whereas with regression problems, Y is continuous (Bramer, 2007). This section provides more information on this method and its application in classifying patients in order to determine rules for readmission.

3.7.1 Classification trees

A classification tree involves the process of asking a certain sequence of questions, where the type of question at each step depends on the previous question's answer. The sequence stops when a class is predicted. The key to tree classification lies in determining how the learning set should be used to 'find good splits and know when to stop' (Breiman *et al.*, 1993). The fundamental idea of splitting is to select a split that results in each child node being 'purer' than the parent subset. A node's purity refers to the number of classes contained in a node. The purer the node is, the more data it contains of only one class (Bramer, 2007).

When a patient who has suffered from a heart attack is admitted to a certain hospital, nineteen variables such as age and blood pressure are recorded during the initial 24 hours. CART can be used to develop a method in order to identify a high-risk patient – someone who will not survive more than 30 days (Breiman *et al.*, 1993). The variables (measurements) must be arranged in a preselected order, e.g. v_1, v_2 where v_1 is blood pressure and v_2 is age. The measurements (v_1,

v_2...) for a certain case is defined as the measurement vector $\mathbf{v}$. The measurement space V is again defined as containing all the possible vectors (Breiman *et al.*, 1993).

The number of cases fall into J number of classes. C can be defined as a set of classes, C=$\{1, ..., J\}$. The class can systematically be predicted by a rule that assigns class membership in C to every measurement vector $\mathbf{v}$ in V. This leads to one of Breiman *et al.* (1993)'s definitions saying, 'a classifier or classification rule is a function d($\mathbf{v}$) defined on V, so that for every $\mathbf{v}$, d($\mathbf{v}$) is equal to one of the classes $1, ...J$.'

With systematic classifier construction, the learning sample contains historical data on N number of cases along with their real classification. A learning sample thus consists of data (v_1,j_1) to (v_N,J_N). Breiman *et al.* (1993) introduces two variable types. The first is a real number, named an ordered or numerical variable, and the second is a categorical variable which can only be one of a finite amount of values, without a natural order. The basic purpose of a classification study can be to determine an accurate classifier or be used to determine the predictive structure of a problem, or both. Determining the predictive structure involves the identification of the types of conditions that determine when and why an object is in a certain class (Breiman *et al.*, 1993).

The starting point of a tree is referred to as the root node and contains all the data entries of the learning set ($\mathcal{L}$). A node is a subset of variables. A parent or non-terminal node splits into two child nodes (binary split) which is determined by a boolean condition on the value of a variable. The variable will either satisfy or not satisfy a condition (Izenman, 2008). A node that does not split is known as a terminal node and resembles a class label. Each observation in $\mathcal{L}$ will end up in a terminal node. One class label may exists at more than one terminal nodes (Izenman, 2008).

3.7.1.1 Strategies for splitting

In order to grow a tree, there are four decisions that must be made beforehand, namely:

1. The boolean splitting conditions at each node;
2. The criteria to split a parent node into two daughter nodes;
3. Criteria to stop splitting; and
4. How to assign a class to a terminal node (Izenman, 2008).

The splitting rules depend on whether the variables are ordinal (continuous) or nominal (categorical). For a continuous variable, the number of possible splits at a node is one less than the number of observed values. With a categorical variable, defined by $\mathcal{M}$, consisting of categories $(\ell_1, ..., \ell_M)$, there are generally $2^{M-1} - 1$ possible splits. The total number of splits is the sum of the number of categorical and continuous splits (Izenman, 2008).

To determine the best split for all variables, the best split for a variable should first be determined. Thus, the goodness of a split must be defined and measured. The algorithm for the decision tree has to make a decision at each node on which variable it would be best to make a split. The split is made to maximise the improvement in the model's prediction accuracy and it is determined by a node impurity measure. Purity refers to nodes containing a larger proportion

of a class – thus, maximising the accuracy or minimising the misclassification error (Kuhn & Johnson, 2013).

With classification, there are mainly two measures, namely the Gini index and cross entropy, which both refer more to purity than accuracy (Breiman *et al.*, 1993). The term accuracy can be misleading as the aim is to partition the data in order to minimise misclassification rather than to partition it so that all the data is classified into mainly one class (Kuhn & Johnson, 2013). Usually, the default method applied by software packages is the Gini index, owing to there being little difference between the two methods for estimating purity (Izenman, 2008). The Gini index is calculated by multiplying the pairs of class proportions for the classes at the node and adding the values together. The equation reaches a maximum when the size of the classes at the node is equal. When the cases in a node are all from the same class the Gini index is equal to zero (Dell, 2015a). More information about the mathematical equations calculating the impurity of a node by either the Gini index or entropy are presented in Section B.2.5 of Appendix B.

Growing a tree thus involves the process of starting at the root node with ($\mathcal{L}$) and then by using the goodness of split criteria for each of the q variables, the algorithm determines the best split at each node for all the variables from X_1 to X_q. The goodness of a split at the parent node is indicated by the reduction in impurity observed by splitting the parent node into its two nodes. The best split at the root node can thus be defined as the node with the largest goodness-of-split function value over all the q best single-variable splits at that node (Izenman, 2008).

The algorithm moves on to splitting the daughter nodes using the same process, but applying the previous split condition (Izenman, 2008). With a binary tree, each parent node is split into two daughter nodes and when the nodes cannot be split further, the tree is saturated. A tree can become very large if it is grown until saturated. There exist a few strategies to restrict the size. Ons such strategy is to establish a condition for when a node is declared terminal, for example if the observations ($n(\tau)$) at node τ is less or equal to a certain value n_{min} (Izenman, 2008). Terminal nodes are not split further and, thus, the higher n_{min}, the smaller the tree will be grown. Another method to reduce tree size is by imposing a minimum value for the goodness-of-split value at a node (Izenman, 2008). These methods are not always advised. A better approach suggested by Breiman *et al.* (1993) is to allow the tree to grow until saturated and then to prune it smaller.

3.7.1.2 Estimating accuracy and tree pruning

As mentioned Breiman *et al.* (1993) believe the tree should be allowed to grow large and then pruned back, which would form a sub-tree of the original larger tree. In some cases, growing it completely results in a complex tree, which is quite similar to the original data set, with nodes possibly containing only one observation. The splitting process can be controlled by allowing the tree to grow until all nodes contain no more than a minimum number of observations or until al terminal nodes are pure. The other method allows a tree to split until a node does not contain more cases than a minimum fraction of sizes of classes. There are various ways to prune a tree where the method is chosen from the estimate of the true misclassification rate (Izenman, 2008).

The pruning algorithm involves growing a tree until nodes contain less than an allocated minimum number of observations and then estimating the resubstitution estimate at each node of

the tree. Afterwards, the tree is pruned toward the root node by minimising the estimate of the true misclassification rate at each stage of pruning (Izenman, 2008). The algorithm constructs a finite sequence of sub-trees decreasing in size. An optimal sub-tree should be chosen and there should be a stopping criterion for pruning. The sub-tree should have the best estimate for the misclassification rate. Breiman *et al.* (1993) suggest using either an independent test sample set or cross-validation. With large data sets, an independent test set is the most efficient and generally preferred. For smaller sets, cross-validation is advised (Izenman, 2008).

The resubstitution estimate $R^{re}(T)$ is generally the number of cases misclassified in a tree divided by the total number of cases that had to be classified. The estimate is however seen as a too optimistic estimate of $R(T)$; sometimes being smaller with bigger trees; and generally grows trees that are too large for the amount of data (Izenman, 2008). The resubstitution error is estimated from the same data that was used for building the tree classifier.

In the case that a test set is used, observations in the dataset are randomly assigned to a learning set and a test set. The test sample is classified by the tree and the true misclassification error can be calculated, because the class of each observation is known. In v-fold cross validation the dataset is randomly divided into v, approximately equal sized, disjoint subsets. The v value is usually taken as either five or ten (Izenman, 2008). An auxiliary tree is built v-times, each time omitting one of the subsets and using it as a test sample. Each subset is used $v-1$ times in the learning set and only once as a test set. The cross validation costs is computed for each of the v test samples and averaged to estimate the cross validation cost (Statistica, 2015a).

3.7.2 Regression trees

Regression trees are in many ways similar to classification and are also known as recursive-partitioning regression. With classification, the class of terminal node required plurality of all observations in that node, where ties are broken randomly. With regression trees, the output variable is given a constant value $Y(\tau)$ at a terminal node τ. The data can be given by $\{(\mathbf{X}_i, Y_i), i = 1, 2...n\}$, where $\mathbf{X}_i$ are observations on an input vector. $\mathbf{X}$ and Y_i observations on a continuous variable Y. The assumption is made that Y is related to X. Furthermore, by using a tree-based method, the aim is to predict Y from the vector $\mathbf{X}$.

The mathematical equations are mostly derived from those used in multiple regression. Pruning a regression tree involves the same methodology as classification trees. It involves allowing a tree to grow into T_{max} by splitting nodes until each node contains less than a predetermined value, i.e. n_{min}. The complexity of the error is measured and sub-trees are formed by means of pruning. The smallest sub-tree of T_{max} is T_k, which is used to estimate the misclassification rate using an independent test set through cross-validation.

3.8 Random forests

Another decision-tree algorithm also developed by Breiman *et al.* (1993) is random forests. The algorithm is a substantial modification of bagging, which is an earlier developed technique used to grow a large group of trees by randomly selecting (without replacement) examples in the training set, and averaging them. Random forests are popular owing to an increased

classification accuracy and being easier to train and modify. The main idea of random forests is to grow many trees (an ensemble or forest) and letting each tree 'vote' for a class (Hastie *et al.*, 2009). The class with the most votes is then selected as the class prediction. The random forest algorithm fundamentally consists of the following steps:

1. For $f = 1$ to F:
   (a) Take a bootstrap[2] sample from the training data of the same size as the data set;
   (b) Grow a random forest tree (RF_f) from the bootstrapped data by repeating steps i, ii and iii for each terminal node, until the minimum node size is reached:
      i. Randomly select h variables from H variables;
      ii. Select the best variable for splitting; and
      iii. Split the node in two child nodes.
2. Output the ensemble of trees $\{RF\}_1^F$ (Hastie *et al.*, 2009).

To make a prediction at a new point, the majority vote of all the random forest trees are taken (Hastie *et al.*, 2009).

3.9 Discriminant analysis

Discriminant analysis is used to determine which variables best predict an instance belonging to two or more groups. For example, a medical scientist may investigate which patient variables discriminate between a patient being likely to recover (i) completely, (ii) partially or (iii) not at all.

Computationally, discriminant function analysis is very similar to ANOVA. If we take a random sample of the heights of 50 males and 50 females, the difference of height will be indicated in the means, with females, on average, not being as tall as males. Thus, height makes it possible to discriminate with chance probability between males and females. Therefore, if a person is short, there is a higher probability of the person being female and vice versa for a tall person.

The concept of discriminate analysis is to predict group membership by examining the mean of a variable to determine whether groups are different. Determining whether two or more groups are significantly different in relation to the mean of a specific variable is similar to a one-way ANOVA problem. If a single variable is tested, an F test will tell whether the variable discriminates between groups. Generally, multiple variables are included in a study in which case a matrix of total variances and covariances and a matrix of grouped intergroup variances and covariances are compared using multivariate F tests. First, statistically significant variables are identified, whereafter it is determined whether the variables have a significant difference in means between the groups (Dell, 2015b).

[2]The basic idea of bootstrapping is to draw data sets randomly, with replacement from the training set, each sample with a size equal to the original training set. This is repeated F times to generate F data sets (Hastie *et al.*, 2009).

3.9.1 Stepwise analysis

In most cases, all variables that are available are included in the analysis to determine which significantly discriminate between groups. Forward stepwise analysis builds the model step-by-step, reviewing variables at each step to determine which has the most discriminating value and including it in the model (Dell, 2015b). Backward stepwise analysis entails including all variables in the model and eliminating the variable which contributes the least, step-by-step. The 'contribution' is determined by the F-to-enter and F-to-remove values, which is an indication of the significance of a variable for discriminating between groups. When employing stepwise analysis, it should be noted that the significance levels is not a true indication of the alpha-error rate (probability of rejecting H_0 when it is true, H_0 is true when there is no discrimination between groups groups) (Dell, 2015b).

3.9.2 Discriminant model for two groups

In the case of only having two groups, the discriminant model is similar to multiple regression. It is also known as Fisher linear discriminant analysis (Dell, 2015b).

With group one and group two, a linear equation can be fitted to the data, where a represents a constant and the b's are regression coefficients for the variables (Dell, 2015b):

$$Group = a + b_1x_1 + b_2x_2 + ... + b_mx_m \tag{3.12}$$

The two-group case follows the logic of multiple regression by determining which variables have the highest contributive strength for predicting group type by selecting the variables which have the largest standardised regression coefficients (Dell, 2015b).

3.9.3 Analysis with more than two groups

In the case where there are more than two groups, more than one discriminant equation, such as (3.12), can be calculated. For example, with three groups, a function can be presented for discriminating between group 1 and groups 2 and 3 combined, and a function for group 2 and 3 (Dell, 2015b).

With multiple groups, groups that are combined in the discriminant functions are determined automatically. The first function has the optimal combination of variables, providing the highest overall discrimination between the groups, decreasing with the second function and so forth. The functions are independent, meaning that the contributions to discrimination between groups do not overlap. This is achieved by canonical correlation analysis, with the maximum number of functions equal to the minimum value of either the number of variables in analysis or the number of groups minus 1. The larger the b standardised coefficient, the larger the contribution of that variable to discriminate between groups (Dell, 2015b).

To obtain an indication of between which groups the various functions discriminate, the means across all the groups for the function are investigated. The factor-structure matrix, which contains correlations between variables and functions, can also be interpreted to derive labels for the functions (Dell, 2015b).

3.10 Selected method

Data mining, as previously stated, is a vast and diverse field, comprising methods that can be applied to the same data set to derive similar results and to achieve equal accuracy. It is not always practical to investigate all alternative methods for one data set. Therefore, the choice of which techniques to employ often depends on the instinct of expert data scientists (Bellazzi & Zupan, 2008). Similarly, the methods that are used in this research are chosen after reviewing the literature gathered on data mining, especially the methods applied to similar published studies, and consulting an expert statistician.

The methods selected for analysing the data set applicable to this project are basic descriptive methods, CART, logistic regression, Cox proportional hazard, discriminant analysis, and random forests. The rationale for selecting the methods along with each goal of analysis are discussed in the following chapter.

3.11 Conclusion: The science of learning from data

This chapter includes an introduction to the field of data mining and statistics. Supervised learning and more popular methods are presented along with more details on methods that are to be used in this research. Studies that have previously evaluated readmission in psychiatric institutions, specifically the methods they used, are also described.

Chapter 4 discusses the process of organising and analysing the real-world data of this research. Similar published studies are also presented with regard to the variables analysed and found significant for predicting readmission. The methodology and methods that will be used for analysing this research's data are also discussed.

CHAPTER 4

Real-world data analysis

Chapter 3 introduced the concept 'learning from data' which entailed an introduction to popular data mining and statistical methods as well as suitable methods for analysing this research's dataset.

This chapter presents the variables investigated in similar published studies whereafter the focus is shifted to the process of putting together the dataset for this research. The data cleaning process is described as well as mention made of any decisions that were taken pertaining to the variables used in the study. Finally, the data learning methodology applicable to this project is presented along with discussion of the various descriptive and predictive methods with application to analysing this project's dataset.

4.1 Variables investigated in similar published studies

Previous studies investigating readmission rates in psychiatric institutions have been discussed in Section 2.3.1 and Section 3.5. Chapter 2 introduced these studies by describing the main objective(s), sample sizes, period of study along with their main findings with regard to readmission. In Chapter 3 the statistical methods used in these studies were introduced. In this section the variables analysed in the studies are discussed and the significant predictors found in each study are indicated.

From the studies displayed in Table 4.1 it seems that most of the studies included all the variables that were available. The most common variables were demographical information, for example, age, gender and employment along with clinical information such as length of stay, diagnosis and history of previous admissions. There are a few exceptions in the studies that are literature reviews or those that focus specifically on a particular population type such as: only males or only depressed patients.

The strongest indicator for readmission varied between the studies, although the common denominator is a history of previous admissions (or readmissions) followed by deinstitutionalisation (shorter length of stay), age and schizophrenia. It is interesting to note that some studies found younger patients more likely to be readmitted where others found this to be the case in older patients. This is difficult to account for owing to interdependency between the variables probably coming into play, for example, a study might have been conducted in a high substance abuse

area and, although it seems that a younger age is a predictor, substance abuse might be the main or a large contributing factor. It is thus important to determine inter-relationships between indicators. Other variables found to be significant by the various studies are gender (specifically male), diagnosis, employment status, substance abuse, treatment non-compliance and marital status. Table 4.1 indicates the variables that were identified as statistically significant indicators for readmission and highlights the strongest predictor variable(s) in each study.

TABLE 4.1: *Variables included in similar studies along with the significant variables.*

Citation	**Variables considered**	**Significant variables**										
		Strongest variable	**# Previous admissions [1]**	**Work status [2]**	**Male [3]**	**Age [4]**	**Marital status [5]**	**Schizophrenia [6]**	**Diagnosis [7]**	**Substance abuse [8]**	**Treatment non-compliance [9]**	**Deinstitutionalisation (LOS) [10]**
(Barekatain *et al.*, 2013)	Education, gender, self-report history of previous admission, age, marital status, type of psychiatric disorder, substance abuse, suicidal behaviour and the length of the current psychiatric disorder.	Type of disorder.							1			
(Bernardo & Forchuk, 2001)	Multiple variables including eduction, marital status, gender, employment, diagnosis, number of admissions, age, and substance abuse.	History of admission.	1		1	younger	1	1				1
(Byrne *et al.*, 2010)	Gender, co-morbidity, marital status, age, socio-economic status, severity of symptoms and LOS.	Diagnosis.	1			1			1			

Variables included in similar studies along with the significant variables, continued.

Citation	Variables considered	Strongest variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
(Durbin *et al.*, 2007)	Systematic review of literature: the readmission period, readmission rate and results of the studies were summarised. All included studies measured discharges from acute care facilities.	History of admissions.	1						1			
(Gillis *et al.*, 1986)	Demographic factors, personal information, social and family variables along with diagnosis.	Various.	1		1		1	1				
(Haywood *et al.*, 1995)	Schizophrenia, unipolar MDD, bipolar disorder and schizo-affective disorders as well as socio-demographic variables such as age, gender, marital status, education and race. Family and housing problems, criminal or violent behaviour, treatment compliance and substance abuse was also included.	Amount of admissions, alcohol and/or substance problems and non-compliance to treatment.	1							1		
(Heggestad, 2001)	Diagnosis, patient turnover, accessibility to therapist, bed-occupancy rate and patient factors (age, gender and marital status).	High patient turnover.										1
(Heslin *et al.*, 2015)	Psychiatric disorders.	Schizophrenia and mood disorders.						1				
(Innes *et al.*, 2015)	Social factors (gender, marital status, age, religion, sexual orientation, ethnicity, income, employment, residence etc.); general health (life satisfaction, general health questionnaire score, co-morbidity index, previous hospitalisation, chronic illness) and lifestyle factors (exercise, diet, smoking and alcohol consumption) were considered.	Age and previous admissions.	1			1						

Variables included in similar studies along with the significant variables, continued.

Citation	Variables considered	Strongest variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
(Johnstone & Zolese, 1999)	Comparing planned short stay with longer length of stay by analysing the length of stay, diagnosis and readmission.	Patients with planned short stay did not experience more readmissions.										
(Jones *et al.*, 2002)	Six diagnostic categories namely adjustment disorders, major depressive disorder (single episode and recurrent), dysthymia, anxiety disorders and not otherwise specified depression.	Major recurrent depression.							1			
(Loch, 2012)	Socio-demographic, psychiatric diagnosis and information about hospitalisation such as LOS, treatment and dosage and psychical restraining. Also, interviews conducted with regard to the patient's general behaviour, follow-up, compliance to treatment, any re-hospitalisations, drug use and the family's opinion.	The family's agreement with permanent hospitalisation of the psychiatric patient.										
(Lyons *et al.*, 1997)	Four dimensions were assessed namely, reasons for admission; complications with the illness (e.g. substance abuse, family disruption); treatment complications; and the severity of the illness.	Self-care impairment and severity and persistence of illness.							1	1		
(Malesu)	Multiple factors including gender, employment status and other demographic characteristics.	Gender and length of stay (schizophrenia).		1	1	1		1				1
(Mark *et al.*, 2006)	Geographic region, gender, race, place of service, age, mean length of stay and treatment follow-ups.	Male gender and alcohol abuse.			1	older				1		
Mayoral *et al.* (2012)	Social and demographical information, number of previous admissions, clinical condition and functioning at discharge, and anti-psychotic treatment at discharge.	Amount of previous admissions.	1	1								

Variables included in similar studies along with the significant variables, continued.

Citation	Variables considered	Strongest variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
(Moss *et al.*, 2014)	Age at admission, schizophrenia and related disorders, gender, education, marital status, LOS, substance abuse, Global Assessment of Functioning at time of discharge, alcohol consumption, employment, income insurance assistance, history of emergency room visits, employment, history of violence, number of previous admissions in two years and psychiatric admissions in the past two years.	Amount of previous admissions.	1			1						1
(Niehaus *et al.*, 2008)	Length of stay, crisis-discharge, income level and marital status.	Crisis-discharge policies.										1
(Wickizer & Lessler, 1998)	Length of stay in comparison with 60-day readmission rates along with characteristics such as age, gender, geographic region, year of review and diagnosis.	Restricted length of stay.										1
(Yussuf *et al.*, 2008)	Socio-demographic variables such as age, gender, marital status, and occupation. Clinical characteristics such as the diagnosis, LOS, number of readmissions, medical history and family history, treatment and compliance to treatment.	Unclear.	1			1		1			1	1
			9	2	4	7	2	5	4	3	1	7

4.2 The Stikland dataset

The management of Stikland Hospital provided anonymised historical data of acute male patients admitted to, and discharged from the hospital. The data capturing and cleaning process was the most time consuming aspect of the project and involved an iterative process over a period of about six months. The initial data gathering process was conducted by the clinical doctor and a data-clerk. During the initial phase review, meetings involving the whole research team were held to discuss the variables that might be available and which would be interesting or necessary to include or additionally capture. The research team constituted the student, study leader, co-study leader and two psychiatrists, Dr. I. Smit and Prof. L. Koen.

The number of variables included in the analysis contributes to the scope of the project and the quality thereof influences the accuracy of the results. This section describes the process of disseminating the data sources, cleaning them and collating all information sources to compile a dataset suitable for data mining and statistical analysis.

4.2.1 The initial dataset

Monthly admission and discharge information was exported from Clinicom in 36 Microsoft Excel workbooks, one for each month of the study period. Clinicom is the medical software suite used by public health care facilities in the Western Cape. The sheets were not all similar in content or format and not all the data fields were complete. A complete monthly data sheet contained eight worksheets of which two sheets respectively included the admission and discharge information. The other sheets were summary data and pivot tables grouped by the various wards. For example: the frequency of the types of admissions for each ward (voluntary, involuntary etc.); transferred from and to locations; frequency of ICD-10 codes; and the number of respective female and male admissions or discharges. As previously mentioned, not all workbooks contained all these sheets, which indicates a lack in consistency with exporting the data from Clinicom. The admission and discharge information available in a complete monthly workbook are displayed in Table 4.2 in the two columns with headings: *Original discharge workbook* and *Original admission workbook*.

The information at first seemed promising and encompassing, with various data fields that may have provided valuable results, however upon further investigation it was noticed that most data fields were empty, partially completed or that not all the sheets had all the information. Obtaining more accurate data was definitely required.

4.2.2 Working towards a complete dataset

The data underwent various data cleaning phases to ensure that data fields were complete and that all the information sources were contained in one dataset. The first step however was to obtain all the required information and complete entries for all the months in the study period.

In preparation for the project, the clinical research partners from Stikland Hospital, further referred to as clinical SMEs, started the process of improving the quality of the data. This process took place from middle August 2015 to 31 March 2016. The process of deciding which

type of data to include was an iterative process involving input from the whole research team. The literature on similar studies led to requesting additional information from the clinical SMEs which included the patient's date of birth, marital status and whether the patient was assigned to a community institution (New Beginnings and/or ACT) or not. The latter was requested after the literature revealed that adequate follow-up and patient programmes did decrease a patient's chance of being readmitted (Niehaus *et al.*, 2008). The marital status could not be captured without individually entering each patient's folder number on Clinicom. This was not feasible at the time owing to it being a time intensive process and it was not expected from the clinical research partners. The patient's area of follow-up, which is the place where the patient receives their prescriptions and follow-up visits after being discharged, was included after being proposed by the SMEs.

Additional data were obtained from ACT and New Beginnings and, after receiving the more complete datasets from Stikland, these were incorporated into one large dataset. During this phase it was found that some of the patients admitted at the community programmes, did not reflect as discharged from Stikland. The clinical SMEs appointed a clerk to re-evaluate and complete the dataset of which 13 months' data were found to be incomplete. Patients admitted in the last few months of 2014 and subsequently only discharged after the study period were followed up on and included in the study until their discharge date in either 2015 or 2016.

Many of the fields for the ICD-10 diagnosis and code were also found to be incomplete. There are four diagnosis fields in the datasets namely: the primary ICD-10 description and code, and the secondary ICD-10 description and code. Some sheets did not contain any diagnostic data and many only had a primary diagnosis and/or code. Based on the published studies presented in Section 4.1, it seemed vital that the data about diagnosis be included in the analysis. After consulting with the clinical SMEs it was decided that the missing diagnoses of readmitted patients would be taken from previous entries, owing to a patient in essence only having one primary diagnosis in their lifetime. This decreased the number of missing values to 560, with 1603 values still missing for secondary diagnosis. The missing entries for the primary diagnosis were finally captured by the clinical SME from one-page discharge summaries kept in the ward which detail each patient's latest discharge. It was decided to exclude the secondary diagnosis and code owing to too many missing data entries.

The data captured from the discharge summaries included the substance use recorded for each admission. The information pertaining to substance abuse was used to compile a second dataset which was analysed separately owing to the initial dataset's information on substance abuse being incomplete. It was decided that the 308 patients for whom there existed accurate substance data were enough for a primary study and that should the results require additional analysis, a data capturer could be appointed to capture the remainder of the patients' information.

4.2.3 Compiling the final dataset

The data required extensive editing and data cleaning to ensure that the information was accurate and in the correct format for analysis with statistical software. All data fields containing check marks had to be converted to either binary or ordinal variables. Recurring patients' various entries had to be combined in one row to ensure that the entries are not analysed as independent events, which would be the case if the information stayed in independent rows.

Additional variables for the subsequent admission(s) were added to the row that contained the first admission's variables. For example, a patient with three admissions had three 'age' variables namely Age1, Age2 and Age3. A macro was compiled in Microsoft Excel to achieve the transferring of information so that each patient only had one entry. This concept is displayed in Figure 4.1 and reduced the number of entries in the dataset from 2361 to 1602.

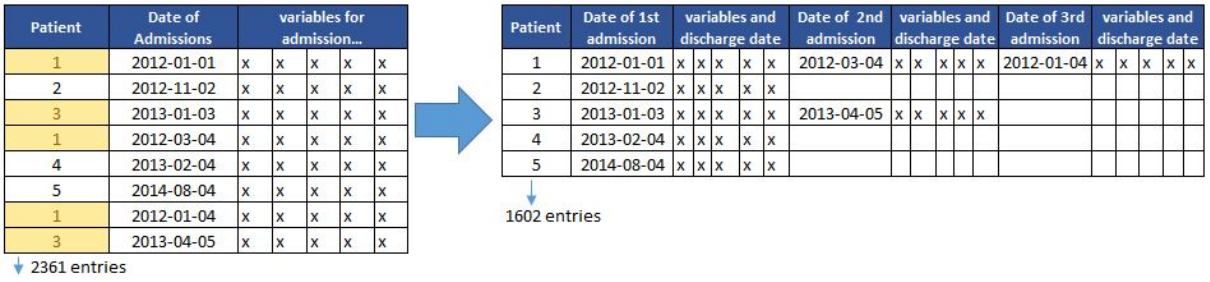


FIGURE 4.1: *Ensuring dependent entries (of one patient) are not analysed independently.*

Table 4.2 displays the information that was contained in the workbooks along with the origin, whether it was additionally captured, recaptured owing to being incomplete but required, or derived from other variables. The variables that are marked as 'inadequate' are not included in the dataset and this is largely as a result of being incomplete, unimportant or not feasible to capture. For example, many variables were not entered correctly onto Clinicom; only reported for a limited period of time; or captured differently by the ward-assistants owing to the definition being unclear. The data indicating whether a patient received 72 hour observation are incomplete, but 'Area 4' (direct admissions) as well as ACT admissions can serve as a proxy for this, as they receive their 72 hour at Stikland Hospital (Koen & Smit, 2016a). The variables that are included in the final dataset are indicated by an asterisk and variables that are not directly analysed, but used to create variables such as age (from date of birth) are also specified. The decision for including a variable or not was made together with the clinical SMEs. The income level is of interest, but is incomplete and regarded as invaluable to analyse. The problem with this variable is not only that the majority of the field are incomplete, but also that it does not always accurately reflect the patients' actual financial situation. For example, involuntary patients are classified as having no income even though they might earn a high income, as they are not charged for the admission (where voluntary patients are charged according to income level) (Koen & Smit, 2016a).

4.2.4 Variables to be tested

There is a total of 2364 data entries which constitutes 1602 patients. Of the total patients, 462 are patients who have been admitted more than once. This is displayed in Figure 4.2. A total of 22 entries were deleted from the original dataset owing to them being incomplete or duplicate entries.

Readmission is generally measured as occurring within either 30 or 90 days. Readmission within 90 days is a global measure for the quality of service given after discharge, but it is arbitrary and one cannot be sure how many admissions a patient may have had for example, at another level 1 hospital before being admitted to Stikland Hospital (Koen & Smit, 2016a). The clinical

TABLE 4.2: *Variables evaluated and included in the analysis.*

Variable	Original admission dataset	Original discharge dataset	Re-captured	Additionally captured	Derived	Inadequate information	Included in final dataset
Admission date$^{[LOS]}_{[Daysdischarged]}$	1						1
Folder number	1	1					1
Area admitted from*	1		1				1
Ward	1	1					
Voluntary admission	1	1				1	
Assisted admission	1	1				1	
Involuntary admission	1	1				1	
State patient	1					1	
Day patient	1	1				1	
72 hour observation	1	1				1	
First admission	1					1	
Readmission	1					1	
Readmission in 90 days	1					1	
Income level	1					1	
Re-admission after conditional discharge	1					1	
Transferred from	1					1	
Discharge date $^{[LOS]}_{[Daysdischarged]}$		1					
Primary ICD-10 description$^{[Diagnosis]}$		1	1				
Primary ICD-10 code		1				1	

Variable	Original admission workbook	Original discharge workbook	Re-captured	Additionally captured	Derived	Inadequate information	Included in final dataset
Secondary ICD-10 description		1	1			1	
Secondary ICD-10 code		1				1	
Crisis-discharge		1				1	
Transferred to	1					1	
Date of birth$^{[Age]}$				1			
Age*					1		1
Follow up*				1			1
New Beginnings and/or ACT*				1			1
Readmission (Y/N)*					1		1
Total readmissions for specific year					1		
Total admissions '12-'14*					1		1
Readmission count*					1		1
Days discharged before readmission*					1		1
Readmission within 30 days*					1		1
Readmission within 90 days*					1		1
Length of stay*					1		1
ICD-10 description*			1	1	1		1
Substance abuse*				1			1

$^{[...]}$Derived from the variable
*Used in analysis

SMEs defined a revolving door patient as one having more than one readmission, and suggested a categorical variable for a patient having either none, one or more than one readmission. Readmission within 30 and 90 day will also be evaluated. A patient's first entry in the dataset will be regarded as the patient's first admission assuming they have never been admitted before, even though this might not be the case. This decision is made owing to not having the previous years' data for the patients. The dependent variable for the data analysis will be whether a patient is readmitted (1) or not (0). The readmission criteria as defined by the clinical SMEs (0,1 and > 1) along with readmission within 30 days (1/0) and readmission within 90 days (1/0)

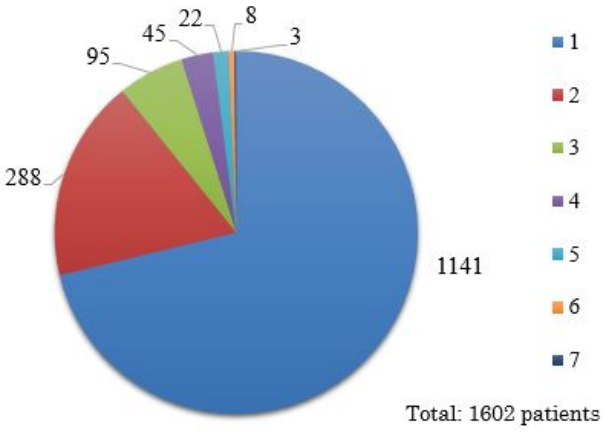


FIGURE 4.2: *Total number of admissions of a patient during the study period.*

will also be investigated. The independent variables are described as being the following [1] :

Admitted from [Area] This categorical variable was reduced from 78 recorded areas to five classes by the psychiatrist and admission nurse:

- [1] Paarl, Vredenburg and surrounds;
- [2] Eerste Rivier Hospital and service-area;
- [3] Karl Bremer Hospital and service-area;
- [4] Direct admissions (Stikland); and
- [5] Other areas of which the majority are not in Stikland's jurisdiction

Follow-up: The place of follow-up was captured as a mixture of codes and descriptions. In the case where a code was unknown, which was the case with only three codes, a sample was drawn from the archive by the clinical SMEs to determine what the code represents. Many codes and descriptions were found to be for the same use. Afterwards there were 19 follow-up possibilities which were further grouped into eight categorical variables namely:

- [PHC] Primary healthcare clinic;
- [Stikland] Stikland Psychiatric Hospital;
- [Tygerberg] Transferred to Tygerberg;
- [Other] Other (long term wards, other provinces, private care, passed away);
- [None] No follow-up;
- [NB] New Beginnings institution;
- [ACT] ACT programme; and
- [NB&ACT] New Beginnings and ACT.

[1] **Dependent variable title**[Variable name in dataset if different than title] description
[class name in dataset] description

Community programme: [Instit] Another categorical variable describing whether a patient was admitted to any community programme or institution:

- [NB] New Beginnings institution;
- [ACT] ACT programme;
- [NB&ACT] New Beginnings and ACT; and
- [None] Do not belong to any programme.

Diagnosis: [ICD10] This categorical variable is a patients primary ICD-10 description or also referred to as a patient's diagnosis. There were 38 categories which were grouped together by the clinical SMEs into seven categories namely:

- [Bipolar] Bipolar disease;
- [GMC] General medical condition;
- [MDD&Anx] Major depressive disorder and anxiety;
- [SA] Schizo-affective;
- [S] Schizophrenia;
- [SIPD] Substance induced psychiatric disorder; and
- [Other] which include attention-deficit hyperactivity disorder (ADHD), adjustment disorder, dementia, intellectual disability, personality disorders, delusional disorder, brief psychotic disorder and intoxication.

Total admissions: This variable is between zero and seven and describes a patient's total number of admissions during the study period.

Days discharged: [Days] This continuous variable describes the number of days a patient was discharged before the current (re)admissions, if applicable.

Length of stay: [LOS] This continuous variable is the length in days that the patient remained in Stikland Hospital after admission and before being discharged.

Age: The patient's age on admission stored as a continuous variable.

Substance: C: A binary (0,1) variable indicating whether a patient is reported using (1) or not using (0) cannabis on admission.

Substance: T: A binary (0,1) variable indicating whether a patient is reported using (1) or not using (0) tik on admission.

Substance: A: A binary (0,1) variable indicating whether a patient is reported using (1) or not using (0) alcohol on admission.

Substance: Other: A binary (0,1) variable indicating whether a patient is reported using (1) or not using (0) substances such as mandrax, cocaine, heroin or opioids on admission.

Substance: None: A binary (0,1) variable indicating whether a patient reported not using (1) or using (0) substances on admission.

Another dataset was created that contained grouped information representing 'lifetime data'. For example, as displayed in Figure 4.3, a patient may have been admitted from Area 2 on his first two admissions, but Area 3 on his third admission, which would then be indicated by a '1' in the column 'Area: 2' and 'Area: 3'. This dataset was more experimental and used for descriptive statistics that may possibly be of interest to the clinical SMEs. Lifetime data for this project refers to the study period (admission(s) between 2012 and 2014).

Patient	Per admission data						Lifetime data									
	FU1	FU2	FU3	Area1	Area2	Area3	FU: PHC	FU: none	FU: other	FU: ACT	FU: ...	Area: 1	Area: 2	Area: 3	Area: 4	Area: 5
1	PHC	PHC	ACT	2	2	3	1	0	0	1		0	1	1	0	0
2	None	ACT		4	4		0	1	0	1		0	0	0	1	0
3	ACT	ACT		3	3		0	0	0	1		0	0	1	0	0
4	PHC			1			1	0	0	0		1	0	0	0	0

FIGURE 4.3: *Example of converting the 'multiple admission entries' to 'lifetime data'.*

A screenshot taken from Microsoft Excel of the initial cleaned dataset, before the entries were grouped per patient, is displayed in Figure 4.4. The yellow cells indicate that a patient (anonymised folder number) occurs more than once. The substance information in green text indicates that the entry's substance information is accurate. From this dataset a macro was used to group the entries of recurring patients in one row. The column for each admission was renamed to clearly indicate to which admission the information was applicable to e.g. Age1, Age2, and so forth. This dataset, where a patients occurs only once, was further developed and used in the various analyses.

Record identifier	Date of admission	# admissions 12-14	Stikland #readmission	Days discharged before readmission	Readmission count	Readmisson	Readmission <30 days	Readmission <90 days	Age	Area admitted from	Date of discharge	LOS	Follow up	ACT/NB/ Both	ICD-10 Coding	Substance	Cannabis	Tik	Alcohol	OTHER (M,H,Cocaine, Opiods etc)	Transferred from [1,0]
1316	2012-08-08	1	0	n.a.	0	1	0	0	22	3	2012-09-25	19	PCH	none	Schizoprehia	none					0
580	2012-08-08	2	>=2	First_Adm	First_Adm	0	0	0	21	4	2012-11-09	22	NB	NB	Schizoprehia	none					0
1135	2012-08-10	1	0	n.a.	0	1	0	0	45	1	2012-09-10	76	PCH	none	SA	none					0
1177	2012-08-10	1	0	n.a.	0	1	0	0	42	2	2012-11-08	22	PCH	none	Schizoprehia	none					0
1264	2012-08-10	1	0	n.a.	0	1	0	0	41	5	2012-08-14	57	PCH	none	SIPD	C,other	1		1		1
1504	2012-08-13	1	0	n.a.	0	0	0	0	31	3	2012-09-06	26	PCH	none	Bipolar	none					1
1542	2012-08-13	1	0	n.a.	0	0	0	0	35	2	2012-08-31	69	PCH	none	MDD and Anxiety	none					1
1014	2012-08-13	2	>=2	First_Adm	First_Adm	1	0	0	45	3	2012-09-11	32	PCH	none	Bipolar	none					0
1472	2012-08-14	1	0	n.a.	0	1	0	0	22	2	2012-10-11	21	PCH	none	Schizoprehia	none					1
788	2012-08-15	1	0	n.a.	0	0	0	0	22	1	2012-10-29	26	STL	none	Schizoprehia	none					1
845	2012-08-15	1	0	n.a.	0	0	0	0	19	4	2012-10-04	21	NB	NB	Schizoprehia	none					1
1227	2012-08-15	1	0	n.a.	0	1	0	0	27	3	2012-09-13	40	STL	none	Schizoprehia	none					1
543	2012-08-15	3	>=2	47	1	1	0	0	23	4	2012-08-30	79	PCH	none	Bipolar	none					0
558	2012-08-15	4	>=2	First_Adm	First_Adm	0	0	0	28	1	2012-09-28	44	NB	NB	Schizoprehia	none					1
1575	2012-08-16	1	0	n.a.	0	1	0	0	19	3	2012-10-26	29	STL	none	SA	none					1
1359	2012-08-16	1	0	n.a.	0	1	0	0	40	2	2012-09-21	31	PCH	none	Bipolar	none					1
606	2012-08-16	6	>=2	First_Adm	First_Adm	1	0	0	44	4	2012-08-29	45	PCH	none	Schizoprehia	none					0
1032	2012-08-17	1	0	n.a.	0	1	0	0	19	5	2012-08-29	53	STL	none	Schizoprehia	none					1
1324	2012-08-17	5	>=2	1	1	0	0	0	36	2	2012-11-05	36	PCH	none	Schizoprehia	none					1
505	2012-08-17	2	>=2	First_Adm	First_Adm	0	0	0	42	2	2012-10-17	38	PCH	none	SA	none					1
1554	2012-08-20	4	>=2	112	1	1	0	0	31	2	2012-09-07	50	PCH	none	Schizoprehia	none					1
1489	2012-08-20	3	>=2	17	1	1	0	0	29	5	2012-09-20	23	PCH	none	Bipolar	none					0
1572	2012-08-20	2	>=2	First_Adm	First_Adm	1	0	0	46	2	2012-09-21	29	TYG	none	Schizoprehia	none					1
942	2012-08-21	4	>=2	First_Adm	First_Adm	0	0	0	45	5	2012-09-27	36	PCH	none	Bipolar	none					0
1009	2012-08-21	2	>=2	First_Adm	First_Adm	1	0	0	40	2	2012-09-14	63	PCH	none	Schizoprehia	none					1
630	2012-08-21	2	>=2	First_Adm	First_Adm	1	0	0	19	1	2012-11-12	61	PCH	none	Schizoprehia	none					1
954	2012-08-22	1	0	n.a.	0	0	0	0	28	1	2012-10-26	18	NB	NB	Schizoprehia	none					1
1020	2012-08-22	5	>=2	36	1	0	0	0	33	3	2012-09-26	35	PCH	none	SIPD	Alcohol			1		0
1206	2012-08-22	2	>=2	167	1	0	0	0	26	1	2012-10-01	36	STL	none	Schizoprehia	none					1
1496	2012-08-23	1	0	n.a.	0	1	0	0	20	2	2012-09-11	63	STL	none	Schizoprehia	none					1
777	2012-08-23	1	0	n.a.	0	0	0	0	35	2	2013-02-04	42	PCH	none	SA	C,other	1			1	1
1162	2012-08-23	1	0	n.a.	0	1	0	0	25	1	2012-09-17	105	PCH	none	SIPD	other				1	0
776	2012-08-24	1	0	n.a.	0	0	0	0	39	2	2012-10-04	39	PCH	none	Schizoprehia	none					0
704	2012-08-24	4	>=2	177	1	1	0	0	20	3	2012-10-22	51	NB	NB	SIPD	Cannibis	1				0
721	2012-08-27	1	0	n.a.	0	1	0	0	19	4	2012-10-22	37	PCH	none	Schizoprehia	none					0

FIGURE 4.4: *Partial screenshot of the dataset in Microsoft Excel.*

4.2.5 Software packages

The software used for analysis is Statistica which was StatSoft's flagship data mining software. The software was purchased from StatSoft by Dell in 2014. The software was chosen based upon referral by a statistical SME and on the fact that Stellenbosch University has an existing license for this software. Statistica has multiple statistical as well as data mining tools which included all of the methods planned to be used for this project.

There are a wide array of open source software programmes available which would also been viable, such as, R-programming, Rapidminer and IMB SPSS to name a few. Statistica has a functionality to integrate with R to enhance or add to its capabilities.

4.3 The data analysis strategy

A key objective of this project is to find variables that are significant indicators for readmission. This will be implemented in the decision making process by identifying and discharging patients who have a lower chance of readmission as determined by the analysis. The relationships investigated along with the methods that are applicable are displayed in Table 4.3.

TABLE 4.3: *Relationships to be analysed and applicable methods.*

Relationship	**Method(s)**
Differences between patients readmitted and not readmitted	Descriptive statistics Chi-square test (nominal data) ANOVA tests (continuous data)
Variables associated with readmission	Logistic regression Decision trees Discriminant analysis
Variables associated with time to readmission	Cox proportional hazard test Kaplan-Meier curves
Additional descriptive relationships and comparisons	Descriptive statistics such as pivot tables, summary statistics (means, frequencies, variance etc.)

The data analysis process is concisely presented in Figure 4.5. The process is iterative at most of the steps and includes regular input from both the research team and statistical SME. The process starts with the dataset where a patient number only occurs once. The dataset is reviewed and organised specifically for the type of analysis. Variables are reviewed to ensure that there are no missing data or patients included that are still admitted, and that the data types are the same per variable (for example no text in a numerical variable).

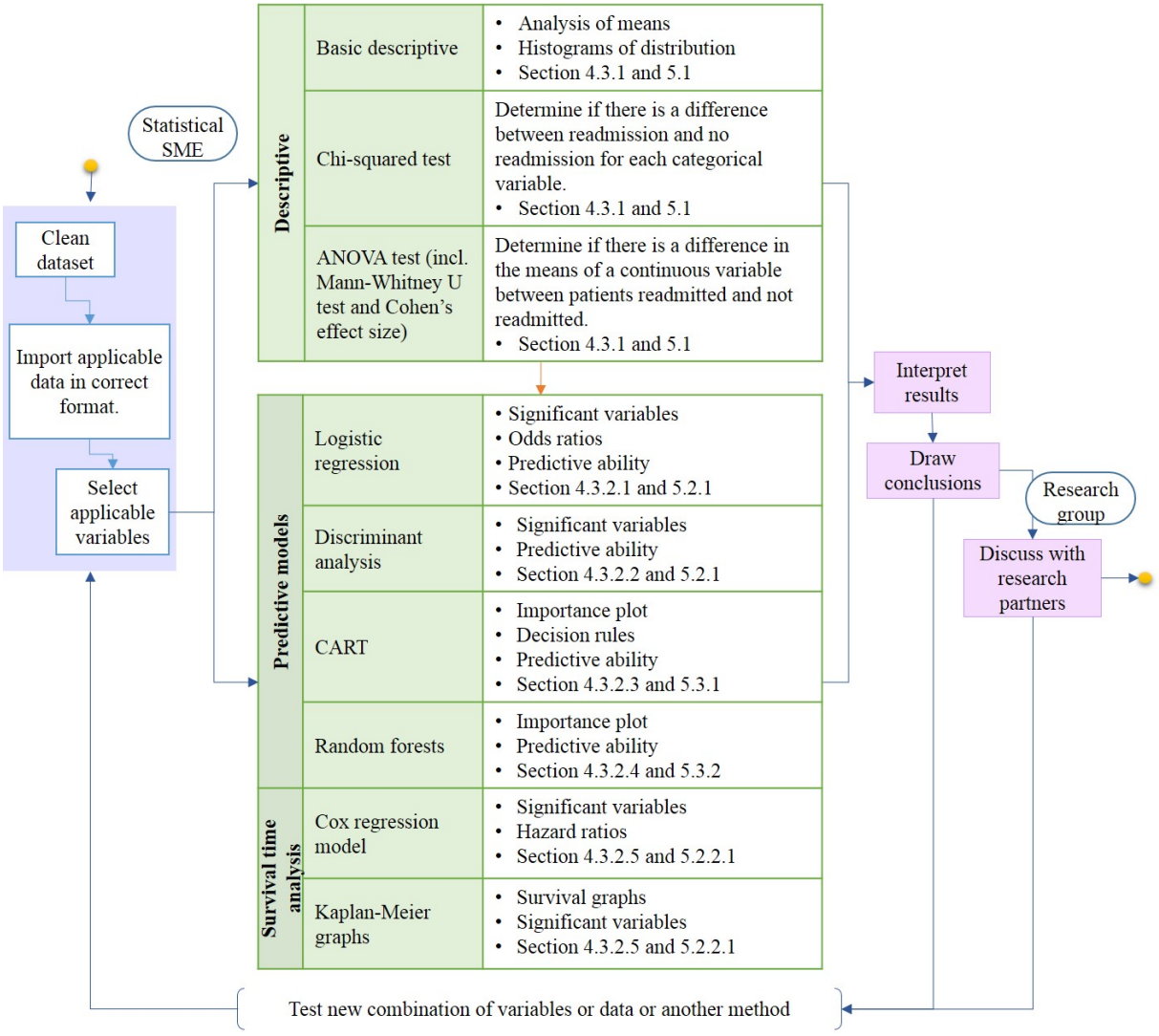


FIGURE 4.5: *Summary of the predominantly iterative data analysis process.*

Categorical variables, such as substance, are split into binary variables owing to the possibility that a patient uses more than one type of substance at an event (admission) as explained in Figure 4.6. All blank values in binary variables should be converted to zeros.

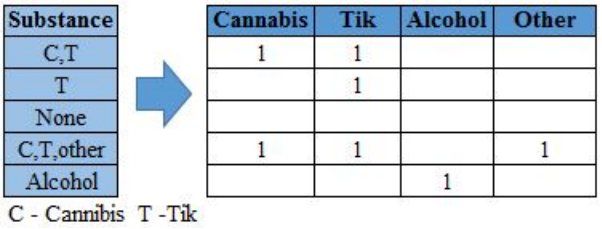


FIGURE 4.6: *Transforming categorical variables into binary variables.*

The following sections aim to describe the statistical and data mining methodologies as imple-

mented for analysing the data pertaining to this project. The methods are explained broadly to minimise duplication when different variables, cases or datasets are analysed with the same method.

4.3.1 Descriptive statistics

Descriptive statistics are the first step in the data analysis process and are performed with Microsoft Excel and Statistica. These statistics merely describe the data and do not aim to predict new data points. The variable classes are compared, with focus given to readmission throughout the various analyses. There are numerous ways in which the variables can be described and compared to each other, but this is not the focus of the project and is not included in order to keep to the point. It is however presented to the clinical SMEs. The results of the descriptive analyses are presented in Section 5.1.

One of the initial analysing steps in Statistica entails drawing histograms of the variables which display the frequency of entries per group in each variable and are valuable for identifying outliers or text in numerical data. Statistica allocates numerical values to text data starting from the number 9999, which is much more than any age or LOS, and thus text such as 'Still admitted' which may occur in the LOS variable, is easily identifiable on the histogram. Classes with few data compared to the other classes in the variable, such as Area 5 in Figure 4.7, can also be identified from the histograms. This is a common phenomenon in medical data. When it occurs merging the groups has to be considered and discussed with both the statistical and clinical SMEs. The risk pertaining to the smaller classes is that a small number of observations are used to draw assumptions that are inferred back to the population. In addition, the rules predictive models derive from few data points may not have the ability to classify the new data correctly (Kidd, 2016a; Izenman, 2008). Another example of a histogram for continuous data is displayed in Figure 4.8, which is the age of the patients on their first admission, indicated by the '1' in the name, *Age1* (and *Area1*).

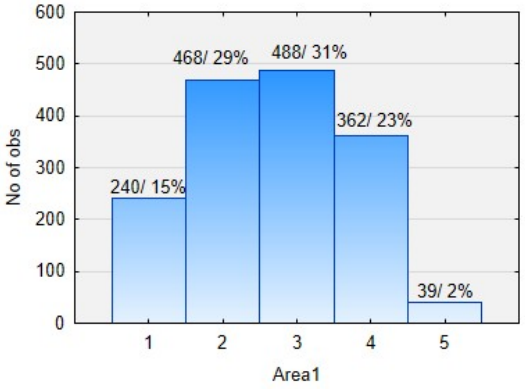


FIGURE 4.7: *Histogram displaying the number of patients admitted from Area 1 - 5.*

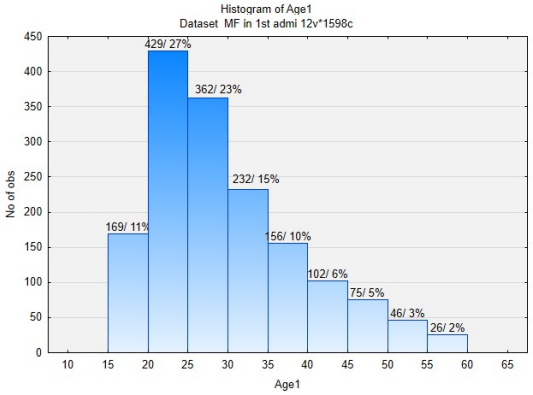


FIGURE 4.8: *Histogram displaying the age of patients.*

To follow, Statistica is used to conduct chi-square tests and analysis of variance (ANOVA). The chi-square test is nonparametric, which implies that there are no assumptions that have to be considered. The test investigates the categorical variables with regard to the dependent variable

which has two classes, namely, readmitted (1) or not readmitted (0). Figure 4.9 and Figure 4.10 display the results of a chi-square test for the area from which a patient originates versus readmission. Figure 4.9 displays a categorised histogram of the distribution of each class (Area 1-5) with regard to readmission. Figure 4.10 displays the chi-square statistic from which it can be determined whether there is a significant difference between readmission (0/1) and the area a patient is admitted from.

In this particular project a p-value of less than 0.05 is regarded as significant owing to it being the most commonly applied. This means that the findings have a five percent probability of not being true. In some cases p-values of less than 0.1 will also be made mention of, which are at a 90% significance level.

The Pearson statistic is interpreted and in this example it can be seen that there is almost a significant difference between patients readmitted and not readmitted with regard to their area. These graphs and tests already convey valuable information about the variables and readmission, but it should be noted that the combinatorial effect of other variables is not taken into account. A table of the descriptive statistics, which contains the same information as displayed in the categorised histograms, is also generated in the output.

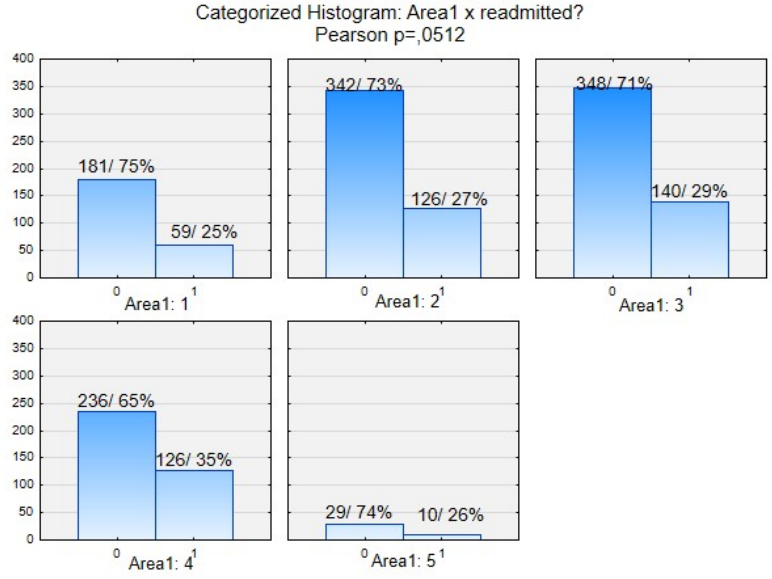


FIGURE 4.9: *Categorised histograms generated by the chi-square analysis displaying the areas from which patients are admitted and if they are readmitted or not (at the first admission).*

	Area (1st ad) x readmitted		
Statistic	Chi-square	df	p
Pearson Chi-square	9,430758	df=4	p=,05119
M-L Chi-square	9,293511	df=4	p=,05417

FIGURE 4.10: *Pearson chi-square statistic for area versus readmission.*

ANOVA is a parametric test used for comparing the average of a continuous variable, such as age, to readmission or no readmission. If there is a difference it implies that age might be an indicator for readmission, which is the same reasoning used with the chi-square test. Two assumptions have to be verified when conducting ANOVA tests: (i) the data is approximately normally distributed, and (ii) the two groups have equal variances. ANOVA is quite robust with respect to normality which means that deviation from this assumption does not have a large effect on the probability of a Type I error (incorrectly accepting a false hypothesis). This is however only applicable if the sample size of one group is equal to or less than 1.5 times the size

of the other group. If the normality assumption is violated, the data can be transformed or the Mann-Whitney U test can be conducted (Laerd Statistics, 2013b). The ANOVA test is run to generate the following output:

Descriptive statistics table which displays the mean, standard deviation, sample sizes and confidence intervals for the parameters;

Levene's test used to evaluate the assumption of homogeneity of variances;

Normal probability plot used to graphically determine if the data is normally distributed; and

Least squares means plot displaying the mean and CI for the continuous variable in both the readmission classes (0/1) along with the p-value of the ANOVA test indicating if there is a significant difference between the two groups with regard to the independent variable.

Normality can be determined numerically or graphically, with the graphical method being used in this project. The numerical test has the advantage of objectivity, but is sometimes overly sensitive with large samples and not sensitive enough with smaller samples. The graphical method is used, with help from the statistical SME with regard to interpreting the graphs correctly (Kidd, 2016a; Laerd Statistics, 2013c).

The larger the sample size, the less effect the deviance from the assumptions have on the test. As the sample size increases (> 40), violation of the normality assumption causes less concern owing to the central limit theorem[2] (Elliott & Woodward, 2007). The dataset for the first admission (both readmitted and not readmitted) can be considered large, as it contains significantly more than 40 samples. Additionally, a significant p-value is generally much more likely with larger samples, which may occur with the dataset (Kidd, 2016b; Field, 2009). Thus, in addition to the Mann-Whitney U and ANOVA test, a macro is included to calculate *Cohen's effect size*, which compares the means of two groups by dividing the difference in means by the average of the standard deviations. A value of 0.2 is regarded as small, 0.5 as medium and 0.8 or more as large (Sullivan & Feinn, 2012).

An example of the normal plot and least squares plot for age on first admission versus readmission is displayed in Figure 4.11 and Figure 4.12 respectively. From the normal plot in Figure 4.12, it can be seen that the ages at first admission are not normally distributed and in the top screenshot (iii) of Figure 4.13 Levene's p-value indicates that variances do not differ significantly and satisfies the assumption of equal variances. Owing to the sample size being large (> 40), violation of the normality assumption is not alarming and the ANOVA results will still be taken into account in conjunction with interpreting the Mann-Whitney U test and Cohen's effect size.

The results of the Mann-Whitney test and descriptive statistics with Cohen's effect size included are displayed in Figure 4.13. From Figure 4.11 the p-value of more than 0.05 indicates that the null hypothesis, which states that the means are equal, is rejected. This indicates that the means of the two groups do not vary significantly. The Mann-Whitney test also supports this by having a p-value larger than 0.05 and similarly, the effect size also found a negligible difference.

[2]The central limit theorem states that sample means are approximately normally distributed for moderately large samples (> 40) even though the population might not be normally distributed (Elliott & Woodward, 2007).

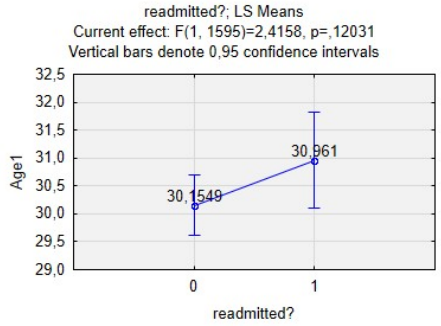


FIGURE 4.11: *Least squares means plot of the age variable versus readmission.*

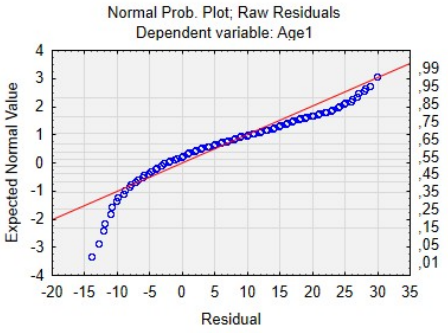


FIGURE 4.12: *Normal plot for the age variable.*

(iii)

	Levene's Test for Homogeneity of Variances Effect: "readmitted?" Degrees of freedom for all F's: 1, 1595			
	MS Effect	MS Error	F	p
Age1	13.5876	31.1336	0.4364	0.5089

(ii)

	Mann-Whitney U Test (w/ continuity correction) By variable readmitted? Marked tests are significant at p <,05000								
Variable	Rank Sum Group1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2
Age1	382447	893556	247740	1.6893	0.0912	1.6908	0.0909	461	1136

(i)

	Descriptive Statistics						
Effect	Level of Factor	N	Mean	Stdev	StdErr	-95%	+95%
Total		1597	30.3876	9.3950	0.2351	29.9265	30.8487
readmitted?	0	1136	30.1549	9.3276	0.2767	29.6119	30.6979
readmitted?	1	461	30.9610	9.5452	0.4446	30.0873	31.8346
Effect size:	0,09(negligible)						

FIGURE 4.13: *Screenshots of the (i) descriptive statistics with Cohen's effect size; (ii) the Mann-Whitney U test; and (iii) Levene's test with regard to age and readmission.*

The Mann-Whitney U test is nonparametric and does not require the assumption of normality. The p-value can be compared to ANOVA's p-value and is interpreted in the same way. The test has a few basic assumptions which this project's dataset satisfies. The first assumption is that the dependent variable should be either ordinal or continuous; in this project it is continuous (years or days). Secondly, the grouping variable (readmitted or not) should be categorical, which it is, and the groups should be independent of each other, which they are. Thirdly, the observations should be independent of each other, thus there should not be a relationship between the observations in each group (a patient only occurs once) (Laerd Statistics, 2015).

4.3.2 Predictive models

To follow, predictive models or classification techniques are explored. The results of the descriptive statistics generally do not influence the data used in this step. In cases where variables contain small groups, it may be decided, along with both the statistical and clinical SMEs, to merge some classes. Some of the previously discussed studies, introduced throughout the previ-

ous chapters, have stated that only variables that seemed to be significant during the univariate analysis were included to develop models further. There is however a risk of excluding variables that may only be found significant in co-existence with other variables. Univariate analysis (ANOVA and chi-square) do not consider co-dependent relationships and effects between the various variables. Logistic regression, discriminant analysis, Cox regression, CART and random forests are multi-variate methods that do test the combined effects of variables. Accordingly, decisions to exclude variables were not made after only considering the univariate results.

4.3.2.1 Logistic regression

Logistic regression is first method to be implemented. This is a popular classification method and is used in many of the studies introduced in Section 3.5. The logit model in Statistica's Generalised Linear Model option is used to build the logistic regression model. The main output that is valuable to this project includes the:

Test of all effects which displays the independent variables included in the analysis along with their (i) degrees of freedom (*dof*), which is the number of classes per variable minus one, except for continuous variables which have a *dof* of 1; (ii) the Wald statistic[3]; and (iii) the p-value. The p-value is the parameter of interest and indicates whether a variable has a significant influence on the dependent variable (readmission);

Odds ratio which is given for the classes of the independent variables and indicates the odds of a patient from a class being readmitted compared to the odds of readmission for a patient not from that class. The p-value is used to identify the odds ratios that are significant. An odds ratio of one conveys no valuable information as it indicates that a patient from a specific class is just as likely to be readmitted as a patient who is not from that class. A ratio of more than one indicates that a patient is that many times more likely to be readmitted and a ratio of less than one, for example 0.5, indicates that a patient from the class is twice as less likely to be readmitted than patients not from that class. When the model is set up, the response variable should be set as 1, which is for a readmission. If it is set to zero, the information will be for a readmission not occurring. The confidence interval (CI) for the odds ratio is also given as a proxy of significance if it does not overlap zero and one (Kidd, 2016c; Szumilas, 2010);

Parameter estimates which are displayed for the independent variable classes and again display the p-value as well as the estimate which is the B_1 coefficient, the standard error and the Wald-statistic. The estimate can be interpreted to get an indication of whether the variable class contributes positively or negatively to the chance of readmission;

Classification matrix which is a table displaying the number of observations correctly and incorrectly classified as either a readmission (1) or not (0) and is used to evaluate the model's predictive classification ability; and

ROC curve which is used to evaluate the model's predictive capability.

[3]The Wald statistic is the outcome of a test which determines if a variable contributes significantly to predicting or estimating the dependent variable by analysing the p-value (UCLA: Statistical Consulting Group, 2011).

The classification ability of the logistic model is represented in a classification matrix displayed in Figure 4.14. Most of the predictive models in this research produce a similar classification matrix. It can be seen from Figure 4.14 that 98% of the no-readmission cases were classified correctly (observed 0, predicted as 0), but only 7% of the readmissions were predicted correctly (observed 1, predicted as 1). Another important method to verify the predictive ability of the model is the ROC curve, which evaluates the goodness-of-fit for a binary classifier. The true positive rate (1 correctly predicted as 1) also known as *sensitivity* is plotted against the false positive rate (predicted 1, when actually 0) also known as the *1-specificity*. The area under the curve (AUC) is a measure of the predictive ability of the model. The closer AUC is to 1, the better the overall ability of the model is to correctly classify cases. The closer the curve follows the Y-axis toward the top (point 1;0), the more accurate the model is and accordingly, the closer the curve is to the 45 degree line, the less accurate the model is (Kidd, 2016a; Statsoft, 2013).

	Classification of cases Odds ratio: 2,733347 Log odds ratio: 1,005527		
	Predicted: 1	Predicted: 0	Percent correct
Observed: 1	32	429	6.9414
Observed: 0	20	1116	98.239

FIGURE 4.14: *Screenshot of the classification table for a logistic regression model.*

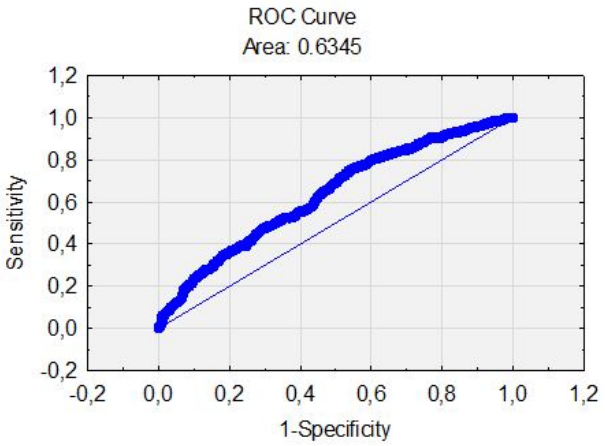


FIGURE 4.15: *Screenshot of a ROC curve also displaying the area under the curve generated by a logistic regression model.*

When analysing the parameter estimates and odds ratios it will be noticed that the last class of the categorical variables is not displayed and thus is difficult to interpret. The variable class that is not included is analysed by running the analysis a second time after changing the order in which the variables are stored in Statistica. The significance and odds ratio for the variable class is then interpreted from the second run.

An indicator that is commonly analysed with regression analysis is the R^2 value and with a binary response variable generalised R^2 values are given namely: the Nagelkerke R^2 and Cox and Snell R^2. Caution should however be exercised when interpreting the values, owing to binary outcome being different to a continuous outcome of simple regression models. The Nagelkerke R^2 value is adapted from the Cox and Snell indicator and thus both can be interpreted although Nagelkerke is preferred owing to the range being between 0 and 1, where Cox and Snell cannot achieve a value of one (Laerd Statistics, 2013a). The proportion of the predictive capability of the model due to factors not included in the model or due to random effects is indicated by 1-R^2 (Riffenburgh, 2012). The recommendation of the statistical SME and literature to rather not interpret the R^2 value as an indication for the predictive capability is accepted and accordingly the goodness of fit and ROC curve will be reported and evaluated.

4.3.2.2 Discriminant analysis

A method similar to logistic regression is discriminant analysis. The *general discriminant* option is selected from Statistica because of its capacity to separately specify the dependent, continuous and categorical variables. The output generated from Statistica is not as extensive as that of logistic regression and consists of the following:

A summary table (multivariate test of significance) displaying all the variables (not the classes) and the p-value which indicates whether a variable has a significant effect on readmission;

A classifciation matrix which is similar to that of logistic regression, displaying the observed 0's categorised as 0 or 1 and the observed 1's either classified as 1 or incorrectly classified as 0; and

A prior probabilities file which is saved as a comma-separated file and is used in R to construct a ROC curve. This is due to Statistica not having the functionality to construct a ROC curve.

4.3.2.3 CART

Next, CART and random forests are implemented to further develop prediction models and determine predictors for readmission. With CART and random forest there are more input settings that can be set or customised which are discussed in this section. Statistica automatically determines the parameters by applying general optimised formulas, but some parameters are customised to the problem. The input settings are determined with help from a statistical SME who is familiar with the project and goals.

Figure 4.16 displays the first tab which is similar to most of the methods discussed and entails selecting the dependent, categorical and continuous variables. The box specifying a categorical response variable is checked and the response variable codes are set as 0 and 1. On the classification tab displayed in Figure 4.17, the misclassification cost, goodness of fit test and prior probabilities are set. The misclassification costs are set as equal, but can be experimented with by specifying a larger cost associated with misclassifying a readmission (classifying a 1 as 0). The goodness fit is set to be evaluated by the Gini index which is the most popular method and also selected as the default by Statistica.

In the case of classification trees the misclassification rate are minimised by minimising the proportion of cases that are incorrectly classified, and is partly influenced by the prior probabilities. If there is approximately an equal number of cases per class, the setting should be set to 'equal' and in the case where the sample is a probability sample, the prior probabilities should be estimated by the sample's class proportions (Statistica, 2015a). For this research the prior probabilities are set as *equal* to it generally offering better results along with selecting *equal* misclassification costs (Kidd, 2016b).

Figure 4.18 displays the tab where the stopping conditions are specified. The tree is set to be pruned based on the *misclassification error* (minimal cost-complexity cross-validation pruning) rather than to be *pruned on deviance* (deviance-complexity cross-validation) or using *FACT-style* pruning. FACT-style direct stopping is selected when the stopping rule is based on a

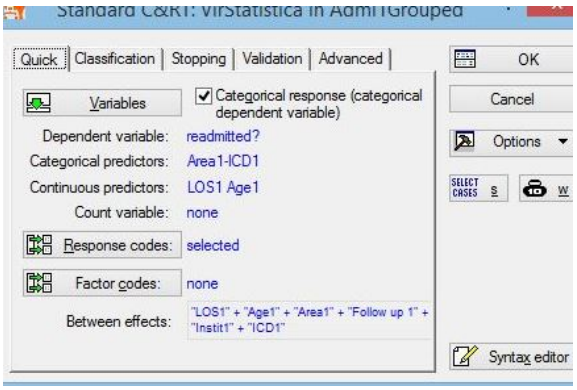


FIGURE 4.16: *Screenshot of the CART input dialog to specify all input variables.*

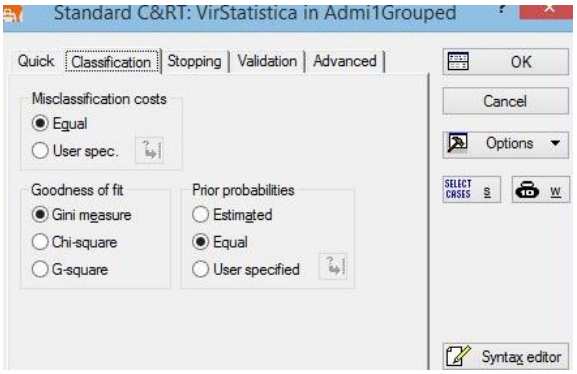


FIGURE 4.17: *Screenshot of the CART input dialog to specify the classification settings.*

required fraction of objects in the node. The only difference between pruning based on the misclassification error or deviance is in the manner that the prediction error is calculated and thus the misclassification error is selected owing to it being more popular and developed by Breiman *et al.* (1993) who is seen as the 'father' of CART (Statistica, 2015a).

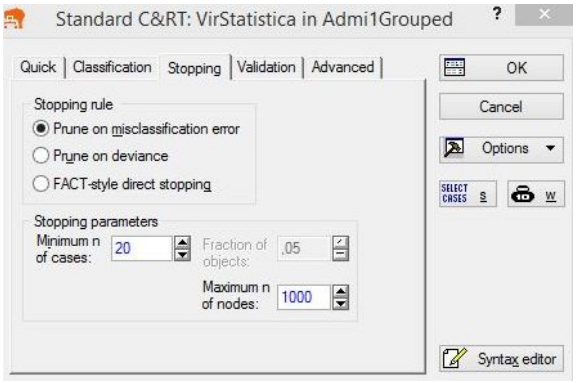


FIGURE 4.18: *Screenshot of the CART input dialog to specify the stopping conditions.*

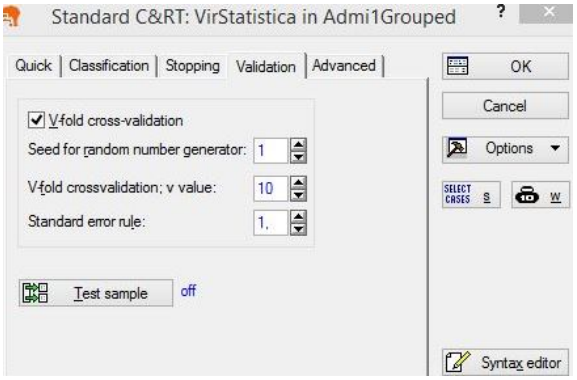


FIGURE 4.19: *Screenshot of the CART input dialog to specify how the prediction is validated.*

The stopping rule is further specified by *minimum n cases* or in FACT-style pruning by *fraction of objects*. It can be specified that the tree should stop splitting after all the cases are classified perfectly, but this may result in a tree that is as complex as the original dataset, with many nodes containing only one observation. This results in a model that may perform poorly in classifying new data. The parameter n for minimum cases is predetermined by Statistica, such as most of the parameters, but is usually chosen a bit smaller, which allows for adequate splitting and growing a large tree (e.g. when Statistica advises 30, 20 is specified). This is acceptable owing to being able to manually validate the best tree selected by Statistica. *Maximum n of nodes* are set to 1000, which is large, but does not affect the tree and is generally changed when the number of nodes should be limited (Kidd, 2016b).

For validation of the tree, displayed in Figure 4.19, v-fold validation is selected owing to the dataset being too small to be segmented in a test set and training set (Refer to section 3.7.1.2).

The cross validation (CV) cost is computed for each of the v test samples and averaged to give the v-fold estimate, as displayed on the cost sequence graph, which is used once again to select the best tree. Minimal cross-validation pruning is achieved if *prune on misclassification error* or *prune on deviance* is selected. The v-value and other parameters are specified by Statistica and used as input for analysis. Nothing is changed on the advanced tab which has only one additional option that can be enabled involving surrogates. This is however not applicable to this research owing to the cases with missing data having been deleted.

After tree pruning, the 'right' sized tree is selected from the set of optimally pruned trees. As briefly mentioned previously, the best tree can be selected manually by analysing the cost-sequence graph, of which an example is displayed in Figure 4.20. The brackets next to the tree number contain the number of terminal nodes (inserted with a macro obtained from the statistical SME). Breiman *et al.* (1993) suggests selecting the least complex tree (least terminal nodes) which has a CV cost close to the minimum. From Figure 4.20 the best tree would be a decision between Tree 50 and Tree 51. Tree 51 will be selected by Statistica as it has the least number of terminal nodes as well as a minimum CV cost, but Tree 50 could also be evaluated and selected for prediction. Tree 48 also has a low CV cost, but is more complex, with seven terminal nodes, which may result in an over-fitted tree which will not be able to classify new data adequately (refer to Section 3.1.3). Over-fitting is also a result of making splits according to randomness or noise in the dataset that was used to build the model and might not be present again in a new set of data (Breiman *et al.*, 1993). The other line which is displayed on the cost-sequence graph is the re-substitution error which was introduced in Section 3.1.3 and which would be more applicable if a test set was used.

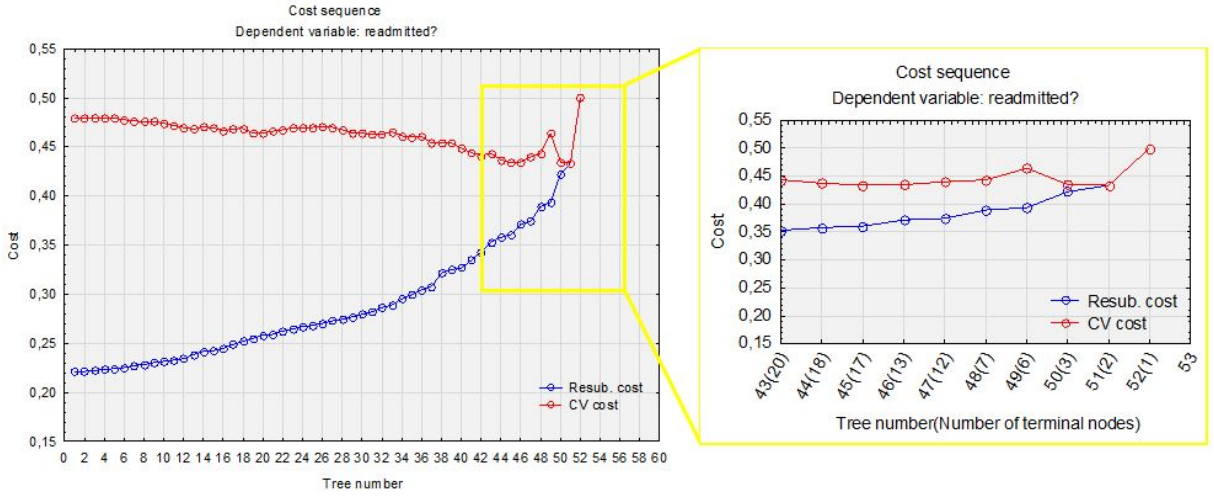


FIGURE 4.20: *An example of the cost sequence graph for all trees in the CART model.*

The output of interest for this project includes the following:

A cost sequence graph which displays the CV cost for the trees and after running a macro, also displays the terminal nodes of each tree which is used to select the best tree(s);

An importance plot which is a bar graph ranking the independent variables according to the importance they play in classifying cases;

A classification matrix Which is similar to the classification matrix of the previous discussed methods, indicating the amount and percentage of cases classified incorrectly and correctly for both readmission (1) and non-readmissions (0); and

A set of categorised histograms additionally generated by a R-macro. An example is displayed in Figure 4.21. The percentage (29%) in the graphs' x-axis title is the probability of a patient being readmitted if selected randomly from the whole dataset. The percentage in the histograms displays the chance of a patient being readmitted (1) if the patient is selected from that group. From the graphs the splitting rules can also be investigated, for example in Figure 4.21 it can be seen that the tree has two terminal nodes and split on the diagnosis. The one rule (left histograms) determines that a patient with SIPD, GMC or another diagnosis have a 16% chance of readmission, compared to 29% if chosen randomly from the dataset.

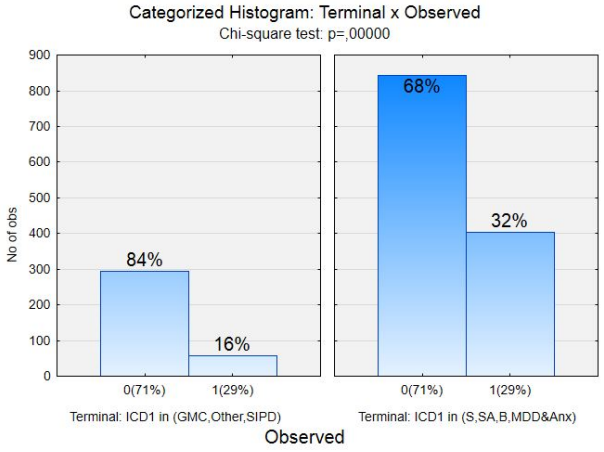


FIGURE 4.21: *Categorised histogram of a selected CART tree.*

4.3.2.4 Random forests

A method similar to CART is random forests which is used to build a predictive model. The output is not as clear as CART with regard to identifying which variables are used for splitting and the distribution of observations in the terminal nodes. This is explained by the algorithm constructing a hundred trees, each classifying a case as either 1 or 0 and thus casting a vote with the majority vote determining the class prediction. The model's performance can however be evaluated and an importance plot is generated, similar to CART.

The variables are selected after which some parameters for the analysis are specified. Figure 4.22 and Figure 4.23 display the user input dialogs for the classification settings and advanced settings respectively. All of the settings are completed by Statistica after the variables for analysis are selected, but some settings can be adjusted to tailor the test in an attempt to get improved classification results. In the classification tab, the misclassification costs can be set as equal or specified. This setting will be experimented with owing to it being possible to set a higher misclassification cost to predicting a readmission as a non-readmission. The prior

probability is set to equal as it generally provides better results, however the effect 'estimated' prior probabilities has on the predictive capability will also be evaluated.

On the advanced tab not much is changed on account of the software using optimal formulas to calculate the parameters. The number of predictors is for example calculated by $log_2(M+1)$ were M is the number of input predictor variables selected. This keeps the number of parameters considered at each node small to minimise the correlation between the trees in the forest and ultimately reduces the error rate (Statsoft, 2015). The random test data proportion is chosen to be as small as possible (0.1) owing to allowing the model to learn from more data points. The subsample proportion is changed to 1, which specifies that the bootstrap samples of the learning set are of equal size, resulting in improved learning. The *minimum n of cases* can be changed to be slightly smaller than specified (not more than 20 in this case) which allows for fewer observations being present in a terminal node, but could also be left unchanged along with the other parameters as determined by Statistica (Kidd, 2016c).

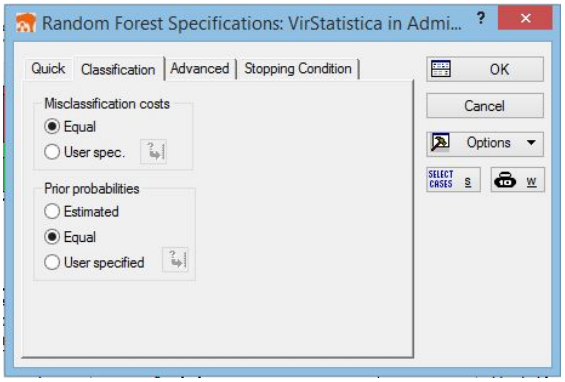


FIGURE 4.22: *Screenshot of the classification tab for generating a random forest model.*

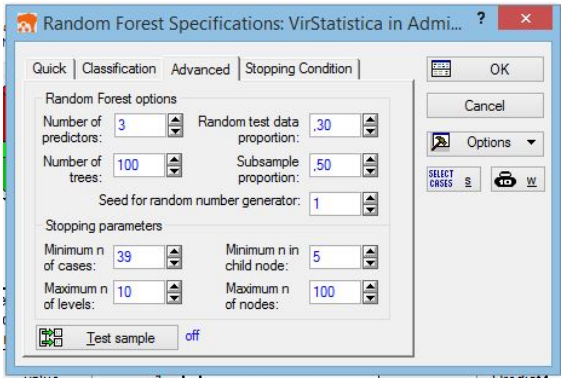


FIGURE 4.23: *Screenshot of the advanced tab for generating a random forest model.*

The output of the random forest model is as follows:

A summary of the forest which graphically depicts the misclassification rate of the test and training set. This may be used to compare forests at various input settings;

A classification matrix for training data which displays the cases (1s and 0s) classified either correctly or incorrectly in the training set which was used to build the model;

A classification matrix for test data which has the same function as the previous classification matrix, but is for the test set which was not used to build the model. The size of the test set is determined in the input settings as a proportion. This gives an idea as to how the model may perform on new data, but unfortunately the test set cannot be chosen as too large on account of limited data points and the model then having to learn from fewer data points; and

An importance plot which is the same as the plot generated by the CART algorithm and ranks the variables according to the predictive role they play in the model. The most important variable is assigned an importance of one. An example is displayed in Figure 4.24.

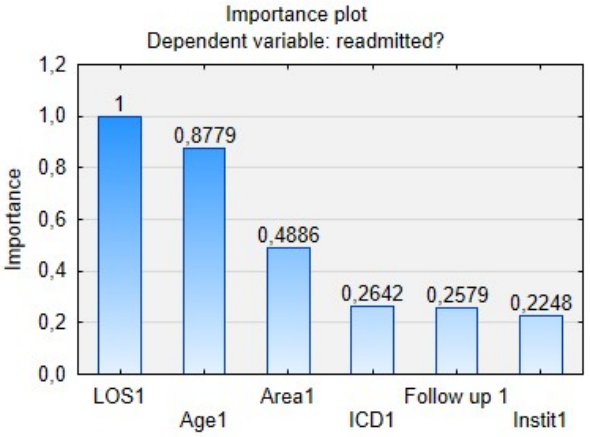


FIGURE 4.24: *Example of an importance plot generated by random forest model as well as CART.*

It should be noted that aside from importance plot no other information is obtained to describe the data or examine readmission and that the model is mainly for developing a prediction model to be used as a decision support tool. The model's predictive capability is evaluated by the classification matrix of both the test and training set.

If the prediction models' predictive capability is adequate, which is relative to the problem under investigation, the code can be used to develop a decision tool. Additional Statistica software packages can be acquired to obtain the code that can be used to build a decision tool. A data analyst or coder should be appointed (for example C++) to alter and incorporate the code into the decision tool, which will be able to allow the probability of readmission as output after the required patient variables are entered as input.

4.3.2.5 Survival time analysis

Cox regression is another method that data of this nature is commonly analysed with, as discussed in Section 3.5 and focus is on the time until readmission. The survival analysis option is selected from Statistica after which regression analysis based on Cox regression is specified. Survival time analysis is not the main aim of the project, but is still relevant and may reveal interesting information to the clinical SMEs.

The input data vary slightly from the other regression models such as logistic regression by adding two variables namely *days survived* and *censored?* as well as transforming the categorical variables in *dummy* variables. The dependent variable is the survival time which is the number amount of days from the first discharge to the next readmission or if not readmitted, until the end of the study period(July 2015). The 'censored' variable indicates whether a patient is censored (1) or not (0). A patient is censored if a readmission did not occur during the study period. The categorical variables' classes are manually transformed into binary variables, for example, the Area variable which consists of five classes (Area 1–5) are transformed into five variables and if the patient originates from Area 2, the Area 2 variable is assigned a 1 and the other four variables a 0. The logistic regression model does this step automatically. Another

step that is manually added is that one class (now a dummy variable) has to be excluded from the model, which is automatically done by logistic regression. To analyse the variable left out, a second model is built by swapping the class out with a class already included. Similarly the same is done for logistic regression, building two models as previously mentioned.

With a Cox regression model (dependent variable: days survived) the independent variables are selected as well as the censoring variable after which a dialog is displayed indicating the general models performance as well as options for other output. The number of censored and uncensored variables are indicated along with the p-value which indicates whether there are variables that have a significant effect on time to readmission (Statistica, 2015c). The output of the Cox regression analysis pertaining to this research are:

Parameter estimates that displays the dependent variables along with the regression coefficient and CI, the standard error, t-value, p-value and risk ratio with its CI. The p-value is calculated by dividing the coefficient with the standard error and is used to determine whether the variable plays a significant role in the time to readmission; and

Survival plots which display survival as a function of the independent variables at their means. A plot can also be drawn for specific variable values or types.

The survival graph can also be customised by supplying input values for the variable, for example, the Area 4 variable can be assigned 0 and then a 1, and used to compare the survival times for patients admitted from the area or not.

The Cox regression model is classified as nonparametric and essentially no assumptions are made about the model and the form of the baseline hazard. There are two factors that should be taken into account when analysing the results, the first being the proportionality assumption which basically assumes that the hazard function, for two observations with different values for the independent variables, does not depend on time. The second assumption assumes that there is a log-linear relationship between the independent variables and hazard function, which means that if a numeric variable increases by one unit it should have the same effect whatever the value of that variable is along with the other variables included in the model (Statistica, 2015c). The Cox regression model is the most common method used for survival data owing to it not assuming any specific survival distribution (Marques de Sá, 2007).

Additionally, Kaplan Meier graphs can be used to individually compare the variable classes with one another based on the proportion of subjects surviving per progressing time.

4.4 Conclusion: Real-world data analysis

This chapter began by presenting the variables that were investigated by similar published studies along with variables they found significant. To follow the original dataset was introduced and the various phases leading to a dataset that can be used for analysis were described. The various descriptive and predictive methods that are to be applied were presented as well as the data learning methodology. Finally, the methods were individually explained with regard to analysis in Statistica and R by discussing the input settings and output concerned with analysing the data pertaining to the research.

CHAPTER 5

Results

Chapter 4 introduced the data 'cleaning process' leading to developing a dataset suitable for analysis pertaining to readmissions at Stikland Hospital. The variables included in the study as well as those investigated in similar published studies were discussed and the methodology and methods to be applied in this project were motivated and explained.

In this chapter the methods discussed in Chapter 4 are implemented, the output is presented and the results are discussed. The chapter starts with describing the data using basic statistical methods, ANOVA and chi-square tests. Furthermore, the indicators for readmission are investigated; survival analysis is conducted; and prediction models are built and evaluated.

5.1 Descriptive statistics

This section aims to describe each of the independent variables pertaining mainly to readmission after the first admission of this study period. Basic statistical indicators are presented along with graphical methods such as box plots and histograms. Each of the independent variables are individually described with regard to readmission.

The admissions after the first admission for the period between 2012 and 2014 are briefly analysed. This is due to having the data and the time to investigate whether additional information could be learnt from the data. The results of the readmissions can be compared to the first admission results as a means of validation as well as to determine whether any trends are present between the subsequent admissions. It is however realised that although referred to as the 'first admission' of this study period, it might not be the first admission of a patient to a psychiatric institution, or, in fact, to Stikland Hospital. The analysis of the variables throughout the study period is presented in Appendix D Section D.1 and in some cases given mention to in this section.

To follow, focus is given exclusively to the data from the first admissions, owing to the amount of data and accordingly the statistical value reducing significantly with each readmission, as displayed in Table 5.1. For example the total number of data points at second admission (first readmission) are the patients readmitted minus those still admitted after their second admission which results in the number of observations reducing from 1597 to 456. The patients still admitted are deleted from the dataset owing to not having a readmission status, follow-up or length of stay for that admission.

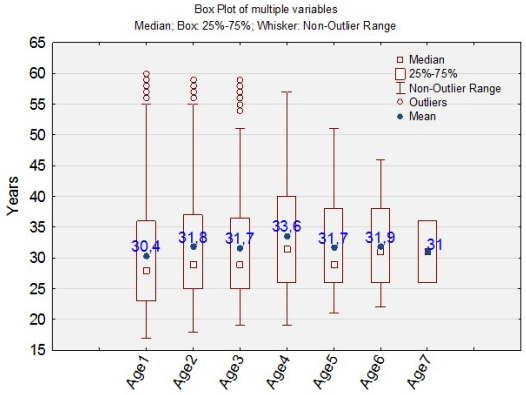
TABLE 5.1: *Size of the datasets reducing significantly at each (re)admission.*

Admission a	Total	# Not readmitted	# Read-mitted	# Still admitted	Total valid data points
Admission 1	1602	1136	461	5	**1597**
Admission 2	461	284	173	4	**457**
Admission 3	173	94	78	1	**172**
Admission 4	78	45	33	0	**78**
Admission 5	33	22	11	0	**33**
Admission 6	11	8	3	0	**11**
Admission 7	3	end of study			

After the first analysis run, it was found that various variables had classes with either very little or many observations which is not ideal for data analysis, especially predictive models. A meeting with the statistical SME and clinical SME was conducted to discuss the statistical complications and review the variables to determine which classes may be grouped. A follow-up meeting was held with both the clinical SMEs to discuss the final grouping of classes. The grouped classes are of the ICD-10 diagnosis and follow-up variables and grouping is discussed in the respective sections describing the variables. Although there is still a large difference between the some of the class sizes, the classes with only a few observations are eliminated. The risk pertaining to the smaller classes is that the information learnt from a small data set may be generalised to describe the whole population. Additionally, the predictive models built from small data points will most probably not have the ability to classify the new data correctly leading to invaluable and incorrect decision advice (Kidd & Smit, 2016; Izenman, 2008).

5.1.1 Age

Box plots of the age of patients are displayed in Figure 5.1 from which it can be seen that the age stays approximately constant throughout admissions, with the average age being between 30 and 33 years. The patients admitted to the acute male wards at Stikland Hospital range between the ages of 18 and 60 which is also reflected in the data. The Clinical SMEs expected a lower average age at the first admission along with an increasing trend between the readmissions, but acknowledge that the first admission, as it is defined in the dataset of this research, might not realistically be the patient's first admission to a psychiatric institution (Koen & Smit, 2016b).



	Descriptive Statistics				
Variable	Valid N	Mean	Min	Max	Std.Dev.
Age1	1597,0	30,4	17,0	60,0	9,4
Age2	457,0	31,8	18,0	59,0	9,5
Age3	172,0	31,7	19,0	59,0	9,1
Age4	78,0	33,6	19,0	57,0	9,1
Age5	33,0	31,7	21,0	51,0	7,7
Age6	11,0	31,9	22,0	46,0	7,1
Age7	3,0	31,0	26,0	36,0	5,0

FIGURE 5.1: *Box plots for patients' age at admission a.*

The average ages of patients readmitted and not readmitted after admission a is summarised in Figure 5.2. The complete least squares means plots along with the normality graphs and Levene's test (to check ANOVA assumptions) are displayed in Appendix C.1. The assumption of equal variances is satisfied, however the normality assumption is violated, but does not cause a concern owing to the sample size being large (> 40) and violation of normality thus not causing concern (Kidd, 2016a; Elliott & Woodward, 2007) (refer to section 4.3.1). The Mann-Whitney U test and Cohen's effect size were also calculated.

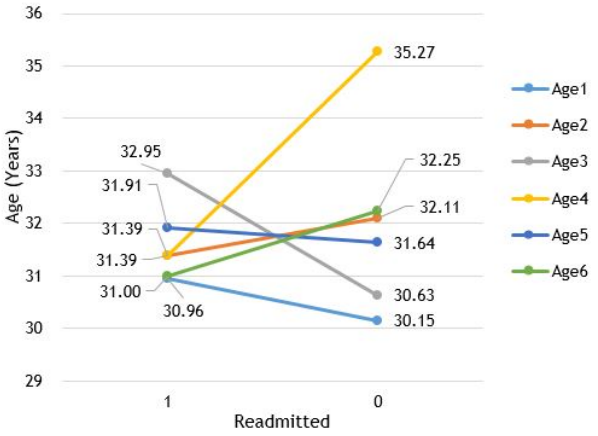


FIGURE 5.2: *Summary of the ages of patients readmitted(1) and not readmitted(0) after admission a.*

A summary of all the p-values are displayed in Table 5.2. It can be concluded that the mean ages between patients readmitted and not readmitted do not vary significantly. According to ANOVA the difference in the average age of patients readmitted after their fourth admission vary significantly at 6% which is confirmed with Cohen's effect size being *medium* along with a *small* difference and significant p-value of the Mann-Whitney test at the third admission. At the third admission the readmitted patients have a higher average age of 33 years compared to 31 years for the non-readmissions. On their fourth admission, the patients readmitted were younger (31 years) than those not readmitted (35 years).

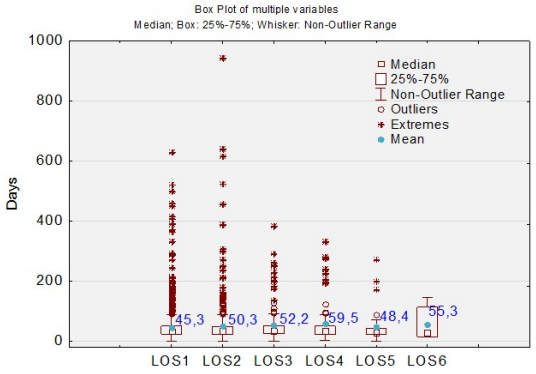
TABLE 5.2: *Summary of the results associated with analysing age and readmission at each admission.*

	ANOVA assumptions		ANOVA	Mann-	Cohen's effect
	Levene's	Normality	p-value	Whitney	size
Age1	0.5777	no	0.1203	0.0912	0.09(negligible)
Age2	0.2853	no	0.4334	0.5411	0.08(negligible)
Age3	0.4204	no	0.0943	0.04615	0.26(small)
Age4	0.1165	no	**0.0620**	**0.0893**	**0.44(medium)**
Age5	0.5269	no	0.9256	0.8786	0.04(negligible)
Age6	0.2465	yes	0.8098	0.9187	0.19(small)

5.1.2 Length of stay

Box plots displaying the length of stay for patients admitted is displayed in Figure 5.3. There is no clear trend with regard to the LOS with most of the observations lying close to the median

and mean (25% to 75%). The average LOS for the first admission is 45 days with 25% to 75% of the data falling between 24 and 50 days and the LOS increasing slightly to an average of 52 days at the third admission. This is similar to the 42 day average length of stay estimated by the clinical SMEs for the acute male ward (Koen, 2016a).



Variable	Descriptive Statistics				
	Valid N	Mean	Min	Max	Std.Dev.
LOS1	1597	45,3	1,0	630,0	48,1
LOS2	457	50,3	1,0	943,0	77,2
LOS3	172	52,2	1,0	383,0	56,2
LOS4	78	59,5	2,0	331,0	73,3
LOS5	33	48,4	1,0	272,0	57,8
LOS6	11	55,3	13,0	146,0	52,0

FIGURE 5.3: *Box plots for LOS during each admission.*

The assumptions for ANOVA were tested and the results are displayed in Appendix C.2 while a summary of the p-values for the various tests is displayed in Table 5.3. The data for admissions one to six are not normally distributed and the Levene's tests were not satisfied by the second, third and fifth admission dates. There is a small difference between the length of stay of patients readmitted and not readmitted at admission two and three.

TABLE 5.3: *Summary of the results associated with analysing LOS and readmission at each admission.*

	ANOVA assumptions		ANOVA	Mann-	Cohen's effect
	Levene's	Normality	p-value	Whitney	size
LOS1	0.855974	no	0.3339	0.0043	0.05 (negligible)
LOS2	0.000107	no	0.00847	0.1158	0.26(small)
LOS3	0.002682	no	0.01869	0.0127	0.37 (small)
LOS4	0.151188	no	0.47879	0.7539	0.17(small)
LOS5	0.039446	no	0.16161	0.3301	**0.55(medium)**
LOS6	0.526161	no	0.7433	0.6098	0.25(small)

On the first readmission (second admission) the patients readmitted thereafter had a much shorter stay (38 days) than those not readmitted (58 days). Similarly the patients readmitted after their third admission had a length of stay of 41 days where those not readmitted had a LOS of 61 days, which can be seen in Figure 5.4. From the analyses it seems that readmitted patients had a shorter LOS than those not readmitted. It might be assumed that the patients with a shorter LOS were crisis-discharges or that this was due to deinstitutionalisation, but there is no evidence in the data to support this claim owing to the crisis-discharge data being inaccurate according to the clinical SMEs, who estimate that the majority of the patients are crisis-discharged (Koen & Smit, 2016a). The assumption will however correspond to previous findings at the hospital which determined crisis-discharge as an indicator for readmission (Niehaus *et al.*, 2008).

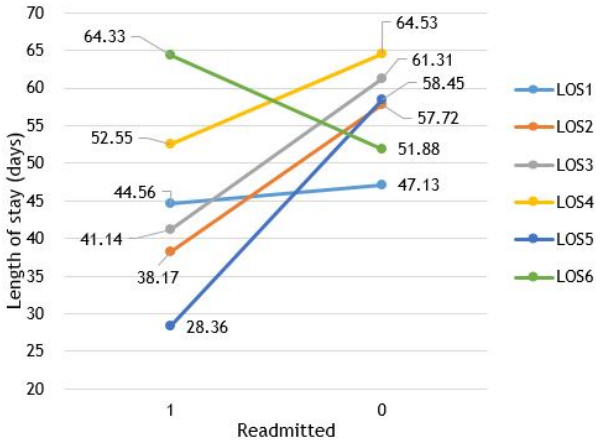
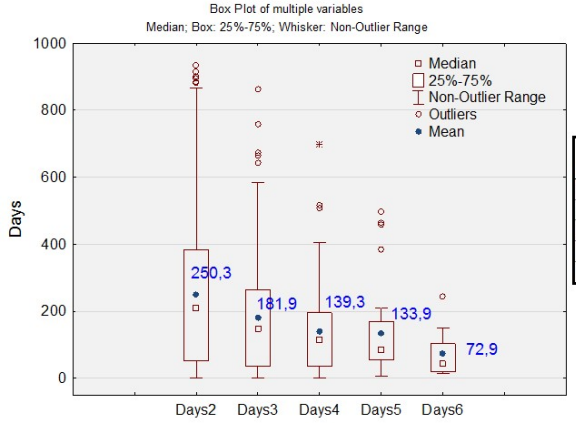


FIGURE 5.4: *Average LOS and readmission after admission a.*

5.1.3 Days discharged before readmission

The days a patient was discharged before a admission were calculated and box plots of the data are displayed in Figure 5.5. From the figure, a slight decreasing trend is noticed from the first readmission (admission two) through to the sixth readmission. The number of days patients are discharged before readmission vary greatly between the patients. This can be seen in the large standard deviation and the 25%–75% percentile being wide.



	Descriptive Statistics				
Variable	Valid N	Mean	Min	Max	Std.Dev.
Days2	457,0	250,3	1,0	933,0	220,4
Days3	172,0	181,9	1,0	864,0	177,6
Days4	78,0	139,3	2,0	697,0	129,1
Days5	33,0	133,9	5,0	497,0	133,3
Days6	11,0	72,9	15,0	244,0	69,8

FIGURE 5.5: *Box plots for the amount of days a patient was discharge before being readmission.*

The mean days a patient was discharged before admission a and whether the patient was readmitted (1) or not (0) after that admission is displayed in Figure 5.6. Additional ANOVA results are displayed in Appendix C.3.

At all the admissions the patients who are again readmitted were previously discharged for a shorter time compared to patients not readmitted. Both ANOVA and the Mann-Whitney U test indicated that patients who were readmitted after their second admission were discharged for a statistically significant shorter time before the second admission compared to the patients not readmitted (average of 215 days versus 272 days). The assumptions for ANOVA were

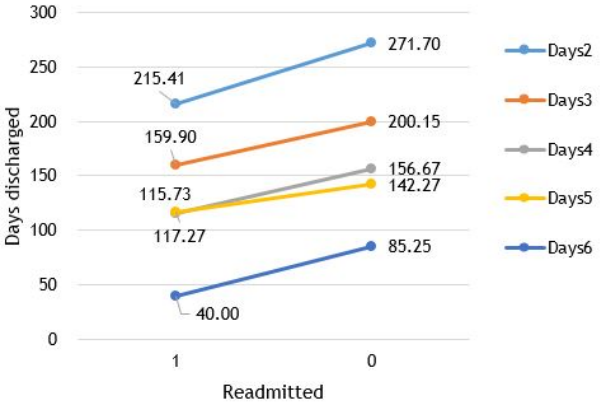


FIGURE 5.6: *Mean days a patient was discharged before admission a and if he was readmitted thereafter.*

not satisfied. Cohen's effect sizes categorised the difference between the groups as 'small'. A summary of the tests' p-values are displayed in Table 5.4.

TABLE 5.4: *Summary of the results associated with analysing the days discharged of patients readmitted or not after admission a.*

	ANOVA Assumptions		ANOVA	Mann-	Cohen's effect
	Levene's	Normality	p-value	Whitney	size
Days2	0.048642	almost	0.0081	0.0147	0.25(small)
Days3	0.023225	no	0.1394	0.4011	0.23(small)
Days4	0.125306	no	0.1680	0.1852	0.32(small)
Days5	0.770234	no	0.6192	0.6062	0.19(small)
Days6	0.115375	no	0.3656	0.6831	**0.71(medium)**

The number of days that patients are discharged for also seems to be decreasing with the number readmissions. The days discharged is not investigated for predictive modelling owing to no data being available on a patient's first admission.

The time to readmission for the patients was also evaluated by investigating readmissions within thirty and ninety days, which is a standard typically measured as an indication of successful discharge planning and outpatient treatment. Readmission within 30 days is typically seen as an indication of inadequate inpatient care, with the first follow-up visit usually occurring on the 30^{th} day after discharge and the patient being given medicine for at least a month. Readmission within ninety days may suggest that a patient did not receive adequate follow-up care for example, ensuring the patient takes their medicine. Figure 5.7 displays the readmissions for admissions one to five. Of the readmissions of the first three admissions, about 20% are within thirty days and about 60% are in more than 90 days which is not alarming to the clinical to the SMEs owing to readmission after 90 days being regarded as much better than readmission within 30 days. Admission within 30 days is considered in response to services during admission, and admission after 90 days is considered in response to services after discharge (Koen, 2016b).

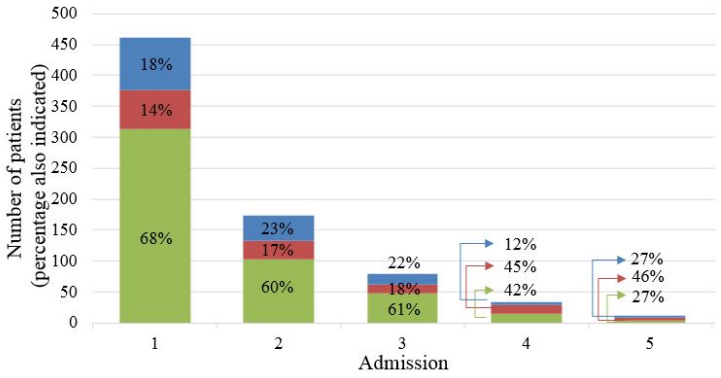


FIGURE 5.7: *Number of patients readmitted within a certain amount of days after an admission.*

5.1.4 Area of admission

As previously described in Section 4.2.4 the area from which a patient is admitted is categorised as either being from (i) Paarl, Vredenburg and surrounds; (ii) Eerste Rivier Hospital and service-area; (iii)Karl Bremer Hospital and service-area; iv) Stikland (direct admission); and (v) other areas (which includes areas outside Stikland area such as Khayelitsha).

A histogram displaying the number of observations in each area at the first admission can be seen in Figure 5.8. Most of the patients are admitted from Area 2 and Area 3 whereafter direct admissions (Area 4), Area 1 and Area 5 follow with very few patients being admitted from Area 5 compared to the other areas.

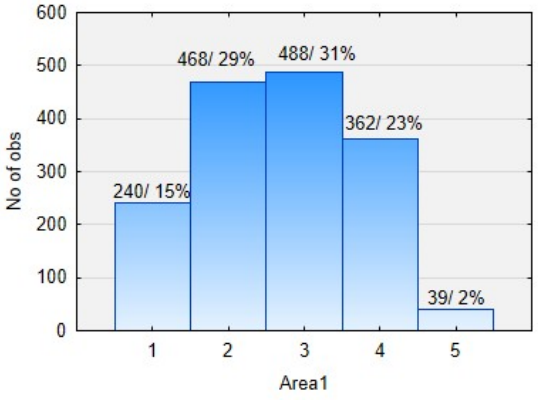


FIGURE 5.8: *Histogram for the areas a patient is from.*

There is no apparent trends throughout the admissions which can be seen in Appendix D.1. During the first three admissions the number of direct admissions (Area 4) increased slightly with patients for Paarl and surrounds (Area 1) decreasing slightly. These trends can be explained by it being more difficult for a patient outside the Stikland area to come back to the hospital directly. This is due to them first having to be examined at a health institution in their area before being referred to Stikland if necessary (Koen & Smit, 2016c).

The histograms from the chi-square test are displayed in Figure 5.9 from which it can be seen

that approximately 74% of patients from Area 1, Area 2, Area 3 and Area 5 are not readmitted and 65% of the direct admissions (Area 4) are not readmitted. It can be expected that more direct admissions are readmitted owing to the patients from the area being easier accessible to Stikland as previously mentioned. This makes it more difficult to interpret results with regard to how Stikland fares against other treatment centres.

The Pearson's coefficient has a p-value of 0.051 indicating that the difference between the readmitted and non-readmitted patients with regard to the area they are admitted from is significant (p=0.051).

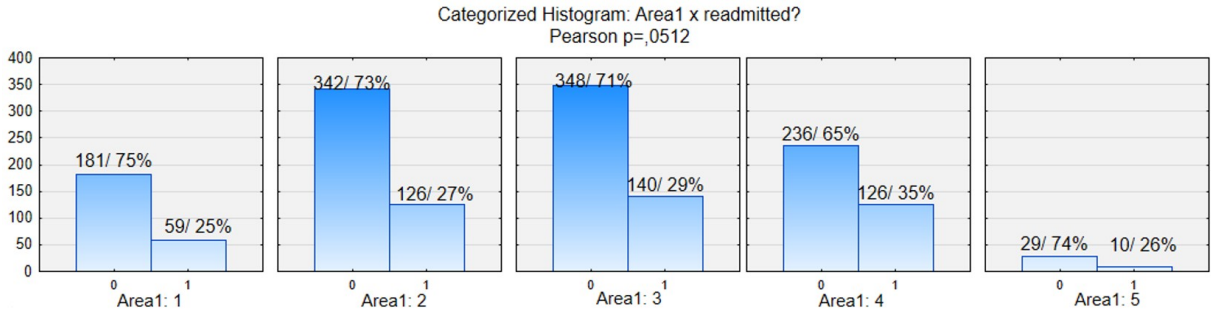


FIGURE 5.9: *Categorised histograms depicting the distribution of the areas grouped by readmission.*

For interest sake, the areas of the second and third admissions were also briefly analysed and are displayed in Appendix D Section D.2.1. The shape of the histograms are more or less the same as that of the first admission with majority of the admissions from Area 3 (32%), Area 4 (28%) and Area 2 (27%)[1]. About 44% of the direct admissions are readmitted in both cases which is an increase from the first admission. The p-values for both the chi-square tests however indicate that there is no significant difference between the patients readmitted and not readmitted with regard to the area at the second and third admission of the research period.

5.1.5 ICD10-diagnosis

Seven groups were initially identified to group the ICD10 diagnoses and the histogram drawn from the first admission data is displayed in Figure 5.10 after which it was grouped as displayed in Figure 5.11. The figures are applicable to the first admission data. The general medical condition (GMC) and 'other' variables were grouped together and schizo-affective, bipolar and 'MDD and anxiety' were grouped together. Schizophrenia and SIPD were kept ungrouped.

The chi-square test found that there is a significant difference between patients not readmitted and those readmitted by only the diagnosis. The categorised histogram and Pearson's p-value are displayed in Figure 5.12. Of the patients diagnosed with schizophrenia 31% were readmitted and similarly 38% of the patients from the schizo-affective, bipolar, MDD and anxiety group (further referred to as SA_Bi_MDD&Anx) were readmitted. Patients with SIPD had a much lower readmission rate of 17%. A similarly lower readmission rate of 16% is observed in the GMC and other group (further referred to as GMC&Other).

[1]The percentages are applicable to the second admission.

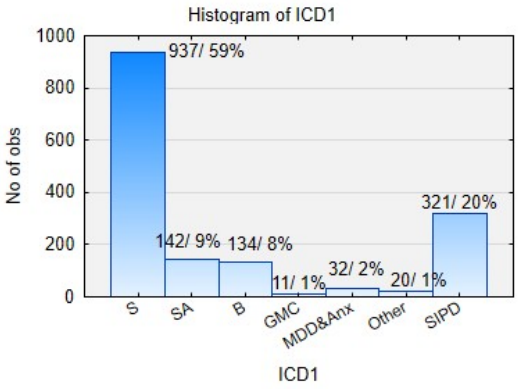


FIGURE 5.10: *Distribution of the ICD-10 diagnoses in the initial dataset for analysis.*

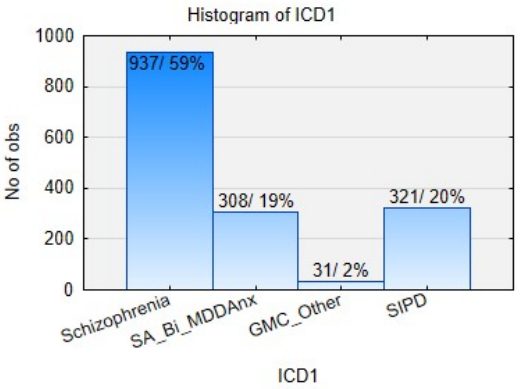


FIGURE 5.11: *Distribution of the ICD-10 diagnoses after grouping.*

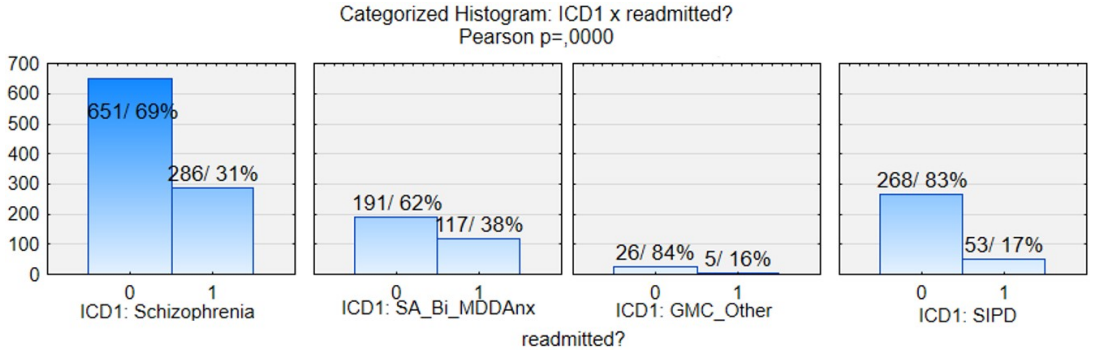


FIGURE 5.12: *Chi-square test results for the 'ICD10-diagnosis' variable versus readmission.*

The clinical SMEs validated the results, but expected a higher readmission rate for schizophrenia. When analysed separately, bipolar disease has a readmission rate of 36% and schizo-affective a rate of 43%. The percentage of patients diagnosed with the various ICD10 diagnoses throughout the various admissions is displayed in Figure D.2 of Appendix D.1. The majority of the male patients admitted at Stikland have a primary diagnosis of schizophrenia followed by SIPD, schizo-affective and bipolar disease.

With the second and third admission data the histograms had a similar distribution as for the first admission data. This can be seen in Appendix D Section D.2.4. The chi-square tests were not significant and the third admission had too little data in the classes to analyse. The second admission data showed similar results to the first admission with regard to schizophrenia and the SA_Bi_MDD&Anx group, but SIPD had a readmission rate of 28% which is higher. The GMC_Other group only had five observations of which all were not readmitted and therefore no valid statistical conclusion could be made.

5.1.6 Follow-up

After a patient is discharged, follow-up consultations are in most cases required, which may take place at either Stikland Hospital, a primary healthcare clinic (PHC), Tygerberg or 'other' which

includes follow-up at correctional services, private hospitals, rehabilitation centres and in other provinces. Another option provided to some patients is to sign up at ACT or New Beginnings. As mentioned in Section 4.3.1 the distribution varied distinctly between the different classes, some consisting of only 14 observations where PHC had over a 1000, which can be seen in Figure 5.13. Accordingly variables were grouped by the clinical SMEs as displayed in Figure 5.14. Patients who received follow-up consultations were grouped according to centre in the following way: Stikland, New Beginnings and ACT were grouped (STL_NB_ACT); and Tygerberg, 'Other' and patients who received no follow-up ('None') were also grouped (Tyg_Other_None). Patients who attended follow-up consultations at a PHC centre were kept separate (Koen & Smit, 2016c).

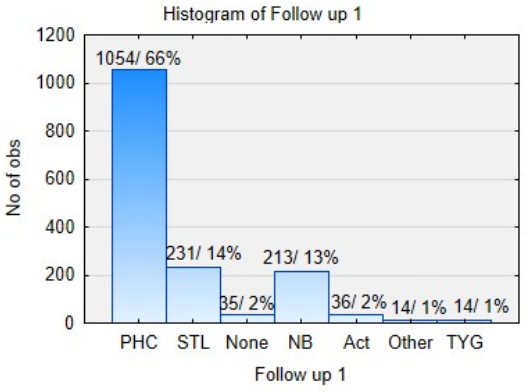


FIGURE 5.13: *Distribution of the place of follow-up in the initial dataset for analysis.*

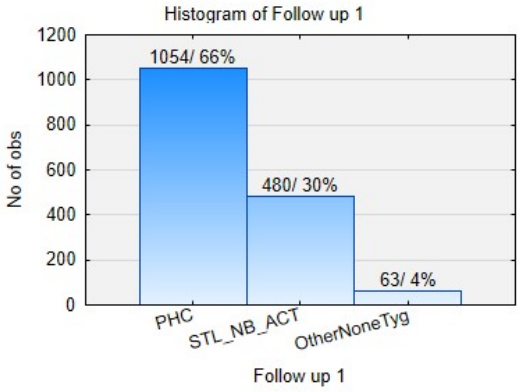


FIGURE 5.14: *Distribution of the place of follow-up after grouping.*

The p-value of the chi-square test displayed in Figure 5.15 indicates that there is no statistically significant difference between patients readmitted or not readmitted with regard to follow-up. From Figure 5.15 it can be seen that approximately 70% of patient in all the classes were not readmitted.

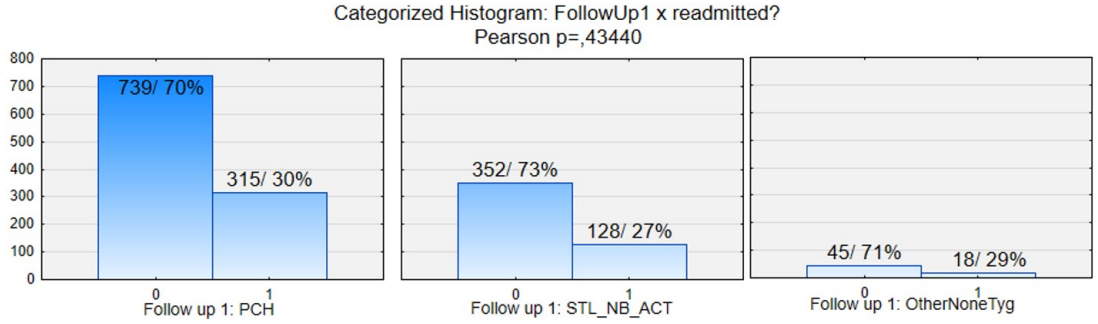


FIGURE 5.15: *Categorised histograms depicting the distribution of the places of follow-up grouped by readmission.*

Follow-up was also summarised in a line graph to determine whether there are any obvious trends between the various admissions, displayed in Figure D.3 of Appendix D.1. The graph does not convey much information from an engineering point of view, however the clinical SMEs considered the case where the place of follow-up at first admission may not have an effect on readmission and only become more important at the readmissions (Koen & Smit, 2016b).

The follow-up for the first readmission (second admission) and third admission were also analysed by means of histograms and the results of the chi-square tests are displayed in Appendix D.2.2. The histograms showed similar results to the first admission. The p-values of neither admissions are significant and the categorised histograms of the third admission have too few data points to make valid conclusions. At the second admission, 35% of the STL_NB_ACT group and 40% of the PHC group are readmitted. This supports the clinical SMEs' opinion that the place of follow-up becomes more important with the number of readmissions. The Tyg_Other_None group has too few data points to analyse.

5.1.7 ACT and New Beginnings

ACT and New Beginnings are psychiatric community care programmes previously described in sections 2.2.7.3 and 2.2.7.4. They are included in the follow-up variable grouped with Stikland, but are also analysed separately on request of the clinical SMEs. For the first admission only four patients simultaneously belonged to New Beginnings and ACT, and accordingly the observations are grouped with the ACT class owing to ACT patients generally also joining New Beginnings, but not the other way around, and also as this is considered to be the input of higher value to future outcomes (Koen & Smit, 2016a). The grouping of classes are displayed in Figure 5.16 and Figure 5.17.

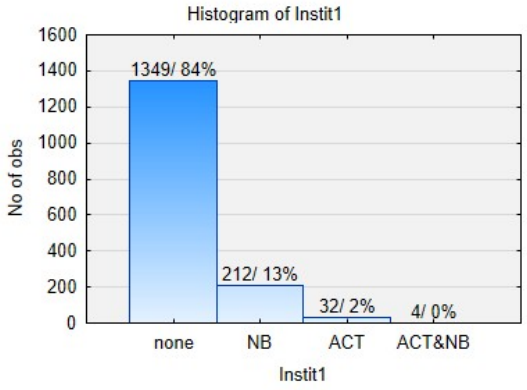


FIGURE 5.16: *Distribution of the community based services in the initial dataset for analysis.*

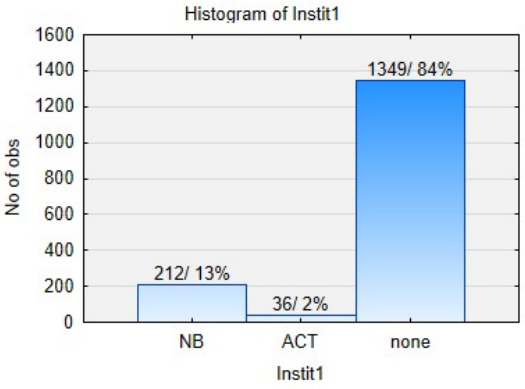


FIGURE 5.17: *Distribution of the community based services after grouping.*

The percentage of patients belong to either ACT, New Beginnings or both throughout the admissions are displayed in Figure D.4 of Appendix D.1. Not much inference can be made from the graph owing to fluctuations possibly caused by patients not being readmitted, patients starting to follow the programmes, or patients changing between the two or being subject to both.

Chi-square analysis was significant with a p-value of 0.006 which can be seen in Figure 5.18. The number of ACT patients readmitted are the same as those not readmitted (the same was found with just the 'NB&ACT'). Of the patients following up at New Beginning 76% were not readmitted, which is similar to the percentage of patients who do not belong to either of the two community care programmes.

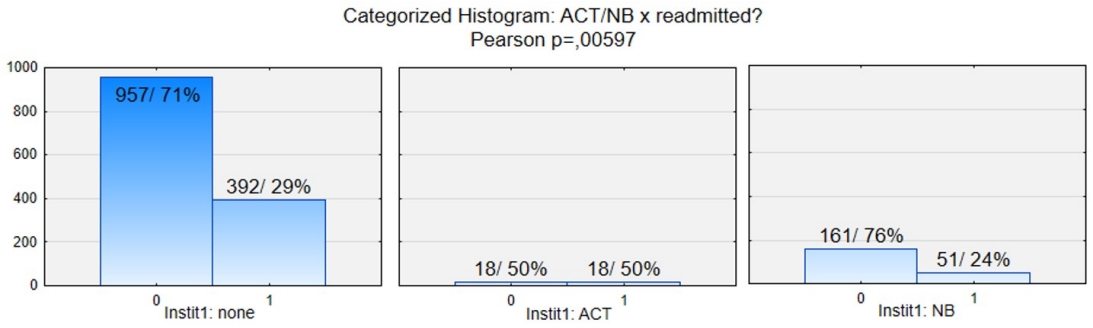


FIGURE 5.18: *Categorised histogram of the ACT/NB variable and readmission.*

For descriptive purposes the second admission (readmission after first admission) was analysed to determine if any significant trend might emerge from the data even though the dataset was smaller. The data of the second readmission was also analysed and is displayed in Appendix D, Section D.2.3. The test for the second admission was also not significant, however it was determined that 38% of the ACT patients and 30% of the NB patients were readmitted. The 'none' group showed an increase in readmissions to 39%.

Throughout the admissions readmission from the New Beginnings group increased slightly, but the reason for this is difficult to determine owing to New Beginnings being a voluntary programme. Patients are therefore able to discharge themselves and not complete the program (Koen & Smit, 2016b). What is interesting to note is that there is not such a sharp decrease in the number of observations in the ACT and New Beginnings classes as that which is observed with the other variables, which may suggest that readmitted patients do start to belong to community care programmes.

5.1.8 Substance use

The Stikland dataset had information on substance use which was inconsistently captured in the secondary ICD10-diagnosis column. As previously mentioned, this column was however incomplete, with some month's data not even having a column for the secondary IDC10 diagnosis. The substance use according to the Stikland dataset is displayed in Figure 5.19 from which it seems that the majority (74%) of the patients did not abuse illegal substance in their admission lifetime for this project.

Figure 5.20 displays the substance use for 307 patients of which the data is sure to be accurate. It can be seen that the substance use profile is opposite to that of the whole dataset displayed in Figure 5.19, with most (76%) of the patients using at least one substance during their various admissions. This confirms the clinical SME's opinion that the data captured in the complete dataset is not accurate. Accordingly, only these 307 patients will be investigated further with regard to substance use and readmissions. The clinical SMEs are additionally interested in specifically methamphetamine (further referred to as tik) abuse owing to their observation that the clinical presentation and response to treatment has become more complex since patients started using tik (Koen & Smit, 2016c).

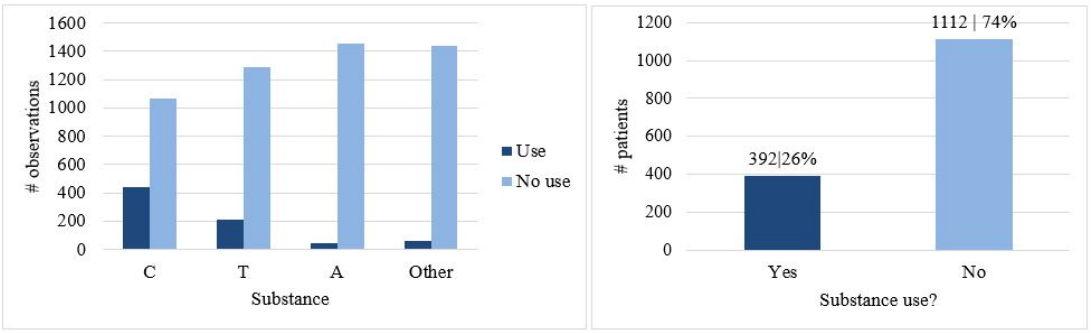


FIGURE 5.19: *Substance use for the whole dataset, suspected to be inaccurate.*

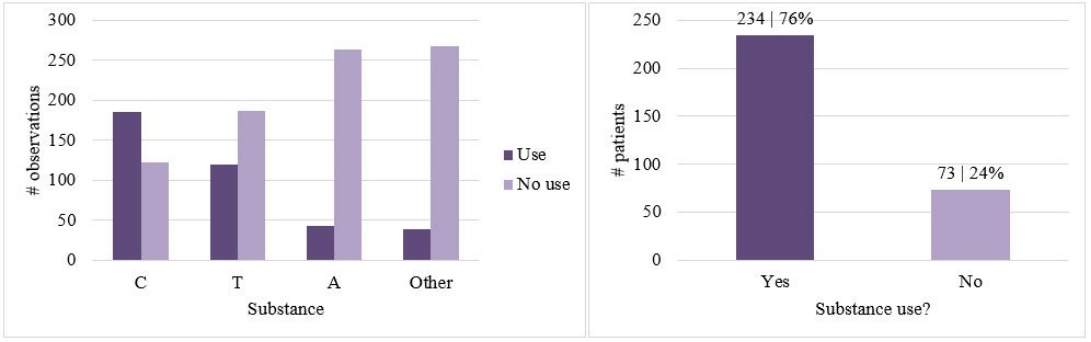


FIGURE 5.20: *A sample reflecting accurate substance use data.*

From Figure 5.20 it can be seen that substance abuse is prevalent among patients with cannabis and tik being the most prevalently abused. Figure 5.21 gives a representation of the simultaneous drug use for the patients. Cannabis and tik co-morbidity is most prevalent and it seems that tik is generally used along with other drugs, especially cannabis. It was found that if a patient predominantly used tik, in 78% of these cases cannabis was also found to be used. Similarly where a patient predominantly used cannabis, 50% of the cases showed evidence of tik use as well.

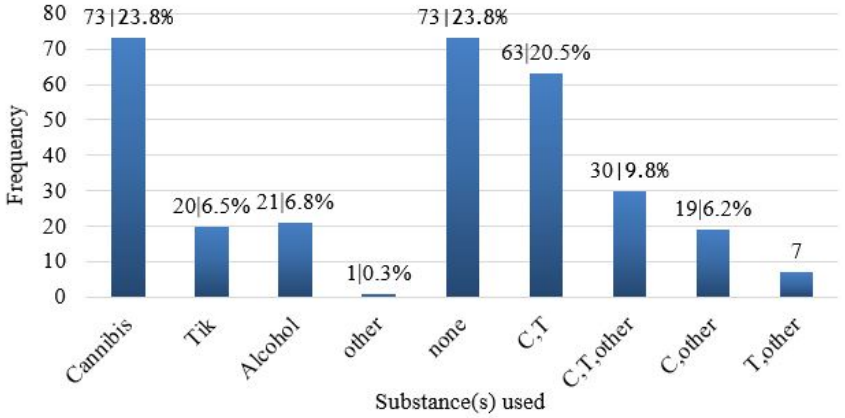


FIGURE 5.21: *Prevalence of multiple substance use.*

The data was further analysed in Statistica using the methods introduced in Chapter 4. Nine patients are recorded as still admitted and had to be deleted from the dataset after the descriptive analysis owing to the text, 'still admitted', in the LOS numerical variable affecting the accuracy of the results. The data of this set contains information of a single patient, but not all the information is from the same admission. Information could relate to a patient's first admission or third readmission and thus it was decided to regard the admission information as 'lifetime'-data (Koen & Smit, 2016a). For example, the age variable contains the ages for the patients at the specific admission that the data is available for. There are many possible analyses to be conducted on the data, but focus will remain on generating information pertaining to readmission at Stikland Hospital.

5.1.8.1 Substance and readmission

Of the patients in the dataset 23% were readmitted after their recorded admission and 24% had admissions before the recorded admission which can be seen in Figure 5.22 and Figure 5.23 respectively. Owing to having little data on readmissions, it is suspected that predictive models built from the substance abuse dataset will not fare well.

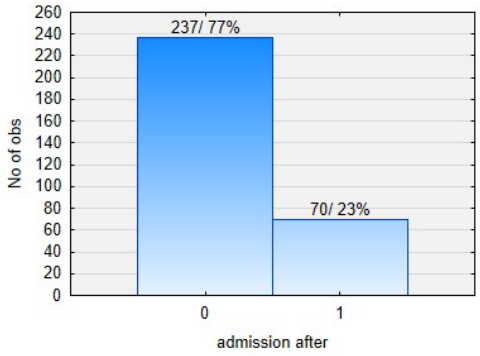


FIGURE 5.22: *Patients readmitted after the recorded admission.*

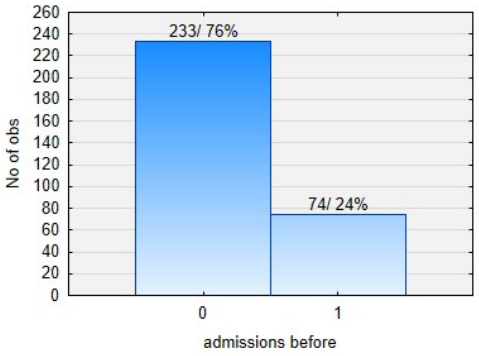


FIGURE 5.23: *Amount of patients with admissions before the recorded admission.*

The data was further investigated pertaining to the substance use which found that 24% of patients using cannabis were readmitted compared to 21% readmitted who did not use cannabis. Similarly, from Figure 5.5, it can be seen that 24% of tik users were readmitted (21% not using tik). Overall 23% of the patients reported to having used substance were readmitted. None of the chi-square test found a significant difference between the use of the specific substance and readmission. The chi-square histograms, descriptive statistics and p-values can be seen in Appendix E Section E.3. From this analysis it seems that substance use is not related to an increased probability of readmission. The dataset is small however, and if more datapoints could be gathered, more definite results may be obtained.

5.1.8.2 Readmission and other variables

The substance dataset was also used to investigate readmission pertaining to the area, follow-up, diagnosis and community care program. No significance between patients readmitted and

TABLE 5.5: *Percentage of patients readmitted who were reported using a substance at admission.*

Substance	Cannabis	Tik	Alcohol	Other	Overall	
Used (Y/N)	Y	Y	Y	Y	Y	N
Readmitted	24%	26%	18%	21%	23%	26%
Frequencies	44/181	30/117	7/40	8/38	52/228	18/70

not readmitted was found in any of the groups and subgroups. This is expected due to the dataset being small with many classes containing few observations. The chi-square histograms are displayed in Section E.3 of Appendix E. The distributions of each subgroup in either the area, follow-up, diagnosis and community care programme variables resemble the bigger Stikland dataset and are not further elaborated on.

As was the case in Section 5.1.1, the age, LOS and days discharged of patients readmitted and not readmitted were analysed to determine if there is a significant difference between the two groups. The output of the tests investigating age, LOS and days discharged can be found in Appendix E. No significant difference between the patients readmitted and not readmitted was found pertaining to age, length of stay or number of days discharged.

5.2 Variables associated with readmission

This section considers whether an increased risk of requiring readmission after a crisis-discharge can be associated with specific variables within the Stikland acute male inpatient population. Indicators for readmission are investigated by applying logistic regression as well as discriminant analysis. Finally, survival time analysis is applied to investigate whether there are variables that significantly influence the time to readmission.

5.2.1 Indicators for readmission

Logistic regression models were built using Statistica to determine whether there are variables that contribute to patients being readmitted and if so, to what extent. Logistic regression provides more descriptive information compared to the other predictive models. Discriminant analysis is also included in this section as it is similar to logistic regression and a good comparison with regard to the predictive capability of the models.

Additional to the grouped dataset, the substance dataset and the 'raw' ungrouped datasets were also individually investigated. The substance data could possibly indicate whether substance use plays a role in readmission whereas the ungrouped dataset is analysed for comparison and interest sake. The ungrouped dataset is however not as statistically and methodologically sound as the grouped dataset owing to it having skewed variable classes. Furthermore, some classes contain very few observations in comparison to other classes as discussed in Section 5.1. The results of the various datasets are critically reviewed at the end of this section. Both the grouped and ungrouped dataset are analysed with regard to the first admissions owing to it containing the most data points as previously explained.

5.2.1.1 Grouped dataset

The following variables: age, LOS, area, follow-up, ACT/NB and the diagnosis from the grouped dataset were included to build a logistic regression model. The area and diagnosis were indicated as significant predictors from the *test of effects output* with the ACT/NB variable being almost significant with a p-value of 0.057. From the *parameter estimates* table: New Beginnings, ACT, Area 4, schizophrenia, SIPD and the SA_Bi_MDD&Anx class were found to be significant. The estimates for New Beginnings and SIPD decreased the likelihood of readmission whilst schizophrenia, SA_Bi_MDD&Anx, ACT and Area 4 increased the chance of readmission.

From the *odds ratios* the following statistically significant observations are made [2]:

1. Patients admitted from Area 4, which are the direct admissions, are 1.7 times more likely to be readmitted than patients from other areas***;
2. Patients diagnosed with schizophrenia are 2.2 times more likely to be readmitted than patients not from the group**;
3. Patients with SIPD are twice as likely not to be readmitted than patients not from this class***;
4. Patients with a diagnosis from the SA_Bi_MDDAnx group are 3.2 times more likely to be readmitted***;
5. Patients belonging to the New Beginnings institution are twice as less likely to be readmitted than patients not belonging to a community program**; and
6. ACT patients are 2.4 times more likely of readmission**.

The ACT odds ratio is alarming owing to it being a community programme, but might be explained by half of the ACT patients being readmitted, and there being only 36 observations in total. The clinical SMEs were consulted about the ACT results and were not too worried, saying that it is possible owing to the ACT programme being structured so as not to eliminate readmission, but reduce the number of readmissions in the long run and also reduce the length of stay of the patients (Koen & Smit, 2016b). In other words, although the rate is high, it is still much lower than it would have been without the ACT input (Koen & Smit, 2016b). The preliminary results of a study conducted in 2008 evaluating ACT patients who belonged to the programme for a period of one year found a reduction in the number of admissions and the length of stay (Botha *et al.*, 2008). In addition, patients who are very ill and prone to be readmitted are subject to this programme. In these cases, from the first admission (of the dataset), readmission is likely to occur. This model classifies only 2.6% of the readmissions correctly, although correctly classified 99% of the non-readmission. The area under the ROC curve is 62% which indicates that the sensitivity and specificity is not adequate.

The discriminant analysis found significant variables to be the area, diagnosis and community programmes which is similar to the logistic regression output. The classification model was built with *equal* prior probabilities and classified 58% of the readmission and 57.8% of the non-readmission correctly. The area under the ROC curve is 0.622 which suggests a poor predictive

[2] $*p < 0.1$; $**p < 0.05$; and $***p < 0.01$

capability. The model built with prior probabilities set to *estimated* classified 99% of the non-readmissions correctly, but only 2.6% of the readmissions. The model is thus not sensitive enough to classify new data.

5.2.1.2 Ungrouped dataset

The output generated by Statistica for this model is displayed in detail in Appendix F, Section F.2.1. The model was built for the readmission-variable to equal 1 which means that the results are in terms of readmission occurring. The community programme variable was omitted for this analysis owing to a warning message appearing indicating that there was too little varying data in the groups.

The initial *test of effects* determined that of the variables included (age, LOS, area, follow-up and ICD10-diagnosis) all variables except age and LOS were found to be significant. Upon further investigation, from the *parameter estimates*, Area 4, schizo-affective diagnosis, bipolar diagnosis, SIPD diagnosis and follow-up at Tygerberg were identified as significant predictors ($p < 0.05$). Follow up at ACT and 'other' were found to be significant at p-values of 0.06 and 0.056 respectively. All of the variables mentioned, especially follow-up at Tygerberg, results in an increase in the likelihood of readmission, except for SIPD in which case the likelihood of readmission is reduced.

From the *odds ratios* the following statistically significant observations are made [3]:

1. Patients following up at Tygerberg are four times more likely to be readmitted***;
2. Schizo-affective patients are 3.8 times more likely to be readmitted than patients who do not have the diagnosis***;
3. Bipolar patients are 2.9 times more likely to be readmitted than patients who do not have the diagnosis**;
4. Patients with schizophrenia are 2.2 times more likely to be readmitted than patients who do not have the diagnosis*; and
5. Patients who follow-up at 'other' are almost 25 times less likely to be readmitted*.

The following odds ratios were also significant at either $p < 0.05$ or $p < 0.1$, but the confidence interval of the odds ratio contains '1' which indicates that there is a chance of having no effect on readmission (refer to Section 4.3.2.1)[4]:

6. Patients directly admitted (Area 4) are 1.5 times more likely to be readmitted***;
7. Patients with SIPD are 1.3 times less likely to be readmitted***;
8. Patients from Area 1 are 1.2 times less likely to be readmitted*; and

[3] $*p < 0.1$; $**p < 0.05$; and $***p < 0.01$
[4] $*p < 0.1$; $**p < 0.05$; and $***p < 0.01$

9. Patients who follow-up at ACT are 2.3 times less likely to be readmitted*.

The predictive capability of the model with regard to readmission seems poor owing to the AUC of the ROC curve being equal to 0.6325. The model is also not sensitive enough which is indicated by the model predicting 98% of the non-readmissions correctly, but only 6.9% of the readmissions (Figure F.6 and Figure F.7).

Discriminant analysis found that the area, follow-up and diagnosis are significant variables for predicting readmission, and support the logistic regression results. A classification model was chosen after evaluating the case where the prior probabilities are assumed to be *equal* and secondly, *estimated* prior probabilities. Both options produced a ROC curve in R with an AUC of 0.636. The classification model with *equal* prior probabilities performed better by classifying 67% of the readmissions and 51% of the non-readmissions correctly. The *estimated* option only classified 7% of the readmissions correctly and 98% of the non-readmissions which is an indication of an insensitive model. The area under the ROC curve also indicates that the predictive capability of the model is poor. The discriminant output is presented in Appendix F Section F.2.2.

5.2.1.3 Substance dataset

The substance dataset was analysed with logistic regression and discriminant analysis to determine if the use of any substance might be a predictor of readmission. The sample is quite small and the data was grouped in the same manner as the 'grouped dataset' to allow for more observation per variable group. From the *test of effects* no variables emerged as significant ($p < 0.05$) although tik use (p=0.836) and ACT/NB (p=0.06) emerged to be significant at $p < 0.1$. Discriminant analysis also found no significant variables with regard to readmission. The output of the model is displayed in Section F.3.1 of Appendix F.

The significant odds ratios' confidence intervals all contain '1' which indicates that the results are not statistically valid (refer to Section 4.3.2.1). This is however expected owing to the dataset being small with some classes containing only a few observations. The results are presented in Section F.3.1 and Table 5.8. The predictive capability is weak and the model insensitive, owing to it predicting all the non-readmissions correctly, but only two out of 70 readmissions correctly. This is supported by the area under the ROC curve being 0.68. The discriminant model classified 68% of the readmissions and 56% of the non-readmissions correctly, which is better than the logistic regression although the ROC curve also had a AUC of 0.68.

5.2.1.4 Comparison of results

From the results generated by the logistic regression and discriminant analyses for each of the three datasets an initial idea can be formed regarding the indicators of readmission. Results similar to those found by logistic regression (Table 5.6) were found by the discriminant analysis (Table 5.7). The ungrouped and grouped dataset both found the area and diagnosis variable to be significant contributors to the models. Additionally the grouped dataset found ACT/NB to be significant at p<0.1 and follow-up was found to be significant in the ungrouped dataset. The

substance dataset did not find any significant variables which may be explained by there being too few observations per group.

	Significant p-values		
	Grouped	Ungrouped	Substance
Area	0.03463	0.03603	
Follow up		0.00738	
ICD10	0.0000	0.0000	
ACT/NB	0.05766		0.06001
Tik use			0.08358

TABLE 5.6: *Significant variables according to logistic analysis.*

	Significant p-values		
	Grouped	Ungrouped	Substance
Area	0.0339	0.0351	
Follow up		0.0140	
ICD10	0.0000	0.0000	0.0906
ACT/NB	0.0443		0.0513
Tik use			0.0866

TABLE 5.7: *Significant variables according to discriminant analysis.*

The variable classes that are significant can be seen in Table 5.8 which also displays the odds ratios in terms of a patient experiencing a readmission (readmitted=1). The confidence interval for the odds ratios is also included in the table owing to it giving a true indication of the significance of the odds ratio. If the confidence interval contains '1' there is a chance that the variable class does not contribute in anyway to the prospect of readmission. It is interesting to note that the ungrouped and substance CIs almost all contain '1' which suggest that the results are untrustworthy and the models' predictive capabilities are inconsistent.

TABLE 5.8: *Odds ratio for variable groups in the logistic regression models.*

Odds ratio for dataset and variable classes					Confidence interval for the odds ratio (-95% — +95%)					
Variable	Variable class	Grouped	Ungrouped	Substance	Grouped		Ungrouped		Substance	
Age										
LOS										
Area	1		0.826*				0.371	1.838		
	2									
	3									
	4	1.703***	1.472***		1.170	2.479	0.680	3.186		
	5									
Follow up	PHC									
	Stikland			3.256*					0.917	11.555
	NB									
	ACT		0.423*				0.112	1.604		
	Tygerberg		3.9657**				1.249	12.596		
	Other		0.0389*	0.307*			0.004	0.413	0.087	1.090
	None									
ICD10	Schizophrenia	2.241**	2.202*		1.605	3.127	1.576	3.076		
	SA		3.838***				2.416	6.099		
	Bipolar	3.195***	2.911**	2.431*	2.169	4.706	1.814	4.671	0.256	23.113
	MDD&Anxiety									
	SIPD	0.446***	0.764**		0.320	0.623	0.243	2.407		
	GMC									
	Other									
ACT/NB	NB	0.407**			0.19	0.85				
	ACT	2.458**			1.17	5.16				
	None			3.392**					0.62	18.51
Tik	No use			0.513*					0.24	1.09

*p<0.1 **p<0.05 ***p<0.01
Red: Increased risk of readmission
Green: Decreased risk of readmission
Yellow fill: contains odds ratio of 1

The following conclusions can be made from the results:

1. Patients admitted from Area 4 are more likely to be readmitted;
2. Patients diagnosed with schizophrenia, schizo-affective or bipolar have a significantly higher chance of readmission. The clinical SMEs validated this by stating that schizophre-

nia is one of the worst diseases. In this case patients suffer from a mood disorder as well as psychosis (hallucinations). SMEs also asserted that schizo-affective patients take a longer period of time to get better;

3. The ungrouped and grouped dataset both indicated that patients with SIPD had a significantly lower chance of readmission, but the ungrouped odds ratio contained '1';

4. The substance dataset did not indicate any significant indicators of readmission, with the exception that not using tik results in halving the chance of readmission compared to using tik. This affirms the opinion of the clinical SMEs who stated that tik has lead to increase the burden on the psychiatric service (Koen & Smit, 2016c). The substance dataset is however questionable owing to it being small and to all the CIs containing '1'; and

5. Patients belonging to New Beginnings are twice less likely to be readmitted whereas ACT patients are more than two times more likely of readmission. This has been previously discussed in Section 5.2.1.1.

In terms of the predictive capability of the logistic regression models, none display adequate predictive capability which can be seen in Table 5.9. All the models could only predict about three percent of the readmission cases correctly, with the grouped dataset predicting 2.6% correct. The grouped dataset is the most statistically sound owing to the classes being more equally sized and containing more observations. The models are effective in classifying non-readmissions, but this is partially explained by there being more observations. For example, if the model classified all the observations as non-readmissions, about 70% of the observations would be classified correctly (all the non-readmissions), but this is not a good model at all in terms of sensitivity. The insensitivity of the models is also reflected in the areas under the ROC curves with 0.6-0.7 being classified as having a poor predictive capability. With regard to the discriminant analysis displayed in Table 5.10, the predictive strength of the models on average seemed better than logistic regression with the grouped dataset having an average correct prediction rate of 59% and an AUC of 0.62.

TABLE 5.9: *Predictive capability of the logistic regression models.*

	Predictive capability		
	Grouped	Ungrouped	Substance
% classified correct			
1	2.60%	6.94 %	2.86%
0	99.03%	98.24%	100%
average	50.82%	52.59%	51.43%
Area under the ROC curve			
area	0.62	0.63	0.68

TABLE 5.10: *Predictive capability of the discriminant models.*

	Predictive capability		
	Grouped	Ungrouped	Substance
% classified correct			
1	58.13%	67.25 %	68.57%
0	57.83%	51.50%	55.70%
average	58.92%	56.04%	58.72%
Area under the ROC curve			
area	0.62	0.63	0.68

5.2.2 Survival time analysis

Survival analysis investigates the variables that may have an influence in the time to readmission. As mentioned in Section 4.3.2.5, the dataset is altered for survival analysis to include the survival time which is the dependent variable, a variable indicating whether an entry is censored or not,

and dummy variables for the classes in each variable. A Cox regression model is built which evaluates whether the variables play a significant role in the time to readmission and Kaplan Meier graphs are constructed to compare survival times for the variable classes.

5.2.2.1 Regression model: Cox regression

With the Cox regression model one class per variable has to be excluded as previously explained in Section 4.3.2.5. Two models are thus constructed with the second run swapping the variable not analysed with one already included. The variable classes included in the first and second model are displayed in Table 5.11. Area 5 was not included in the model owing to it being much smaller compared to the other classes and possibly would have affected the significance inaccurately if included in the model. Area 5 was tested beforehand and had no significant effect on the survival time. The results displaying the significant variables' hazard ratio and confidence interval are also displayed in Table 5.11.

TABLE 5.11: *Variables included in the survival analysis along with the hazard ratio and significance results (built on Cox regression model).*

Variables	First run	Second run	Hazard ratio	CI lower	CI upper
Age	✓	✓			
LOS	✓	✓			
Area: 1	✓				
Area: 2	✓	✓			
Area: 3		✓	1.28*	0.964	1.721
Area: 4	✓	✓	*1.5****	*1.158*	*2.102*
Area: 5					
FU: PHC	✓				
FU: Tyg_Other_None		✓			
FU:STL_NB_ACT	✓	✓			
ICD: S	✓				
ICD: SA_Bi_MDD&Anx	✓	✓	2.63**	1.071	6.466
ICD: GMC_Other		✓			
ICD:SIPD	✓	✓	*0.53****	*0.396*	*0.718*
Instit:None		✓	0.56**	0.3308	0.9641
Instit:ACT	✓		1.80**	1.054	3.069
Instit:NB	✓	✓	*0.53**	*0.306*	*0.909*

*p<0.1 **p<0.05 ***p<0.01
Red: Higher risk of readmission
Green: Lower risk of readmission

The hazard ratio displayed in Table 5.11 is an instantaneous event rate indicating the likeliness that an individual at time t has an event (readmission), assuming the patient is event-free (survived) up to time t. For example, a hazard-ratio of two indicates that patients from the group are likely to have twice as many readmissions (events) in proportion to those not from the group at any time t. The hazard ratios that are displayed in *italics* are variables that were included in both the runs, but only came out as significant in the second. This might be owing to a different group being excluded and with the exclusion, the variable becomes significant. This is not uncommon in medical data owing to groups being unequally sized and it gives an indication that the results should be verified by clinicians (Kidd, 2016c).

Both the models were found to be significant overall in analysing survival times ($p<0.01$). Area 3 is significant at $p<0.1$, but the risk ratio is not statistically valid owing to the CI containing '1'. Patients admitted from Area 4 have a higher occurrence of readmissions and this area plays a significant role in the time to readmission. Area 4 only emerges as significant with the second run. Patients from the schizo-affective, bipolar and MDD and anxiety group have almost two and a halve times more readmissions occurring than patients not from the group. On the other hand, SIPD patients have half the number of readmissions occurring than those not diagnosed with SIPD.

The results of the community institutions ACT and New Beginnings vary with each run and therefore the results are deemed as invalid. It is advised that more data should be captured and analysed before making a conclusion about this group. The ACT group emerges as significant when compared to New Beginnings and 'none', with ACT-patients having more readmissions. With the second run New Beginnings patients were found likely to have half the amount of readmissions as patients not form the group. This makes sense, but the group where patients did not belong to either institutions were also found to have a significant lower amount of readmissions. Thus, no valid conclusion can be made from this variable and its classes.

5.2.2.2 Kaplan-Meier analysis

Kaplan-Meier graphs were constructed for each of the variables and their respective classes and were used to determine whether there is a significant difference between the time to readmission of the different classes. The curves for the Area- and Follow-up variable are respectively displayed in Figure 5.24 and Figure 5.25. No significant difference between the classes and their time to readmission were found.

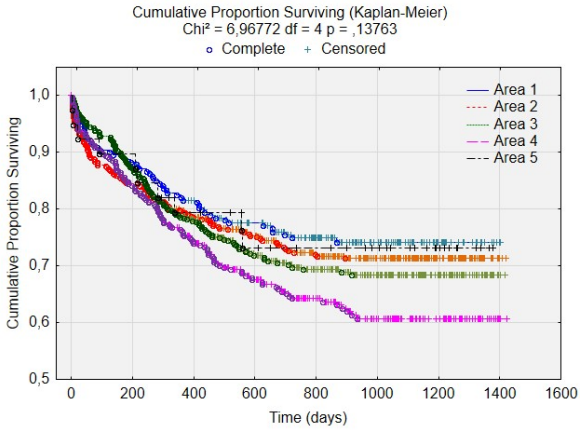


FIGURE 5.24: *Kaplan-Meier survival curve for the area.*

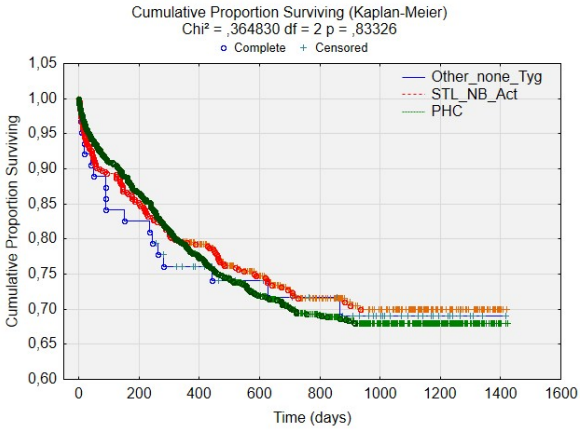


FIGURE 5.25: *Kaplan-Meier survival curve for follow-up.*

The survival time and proportion surviving do vary significantly between the various diagnostic classes as displayed in Figure 5.26. Patients belonging to the schizo-affective, bipolar and MDD and anxiety groups are readmitted in less time when compared to the other groups. In these it was found that fewer SIPD patients are readmitted overall along with this readmission occurring after a longer period of time. The Kaplan-Meier curve for the community programme variable

also found a significant difference between the groups, but the results are analysed with caution owing to previously observing contradictory results. Many ACT patients seem to be readmitted much sooner than those from New Beginnings and 'none' with the proportion of patients not readmitted after about 190 days only being 65%.

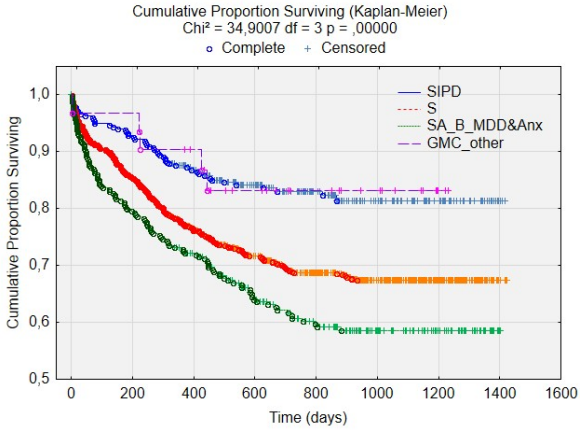


FIGURE 5.26: *Kaplan-Meier survival curve for the diagnoses.*

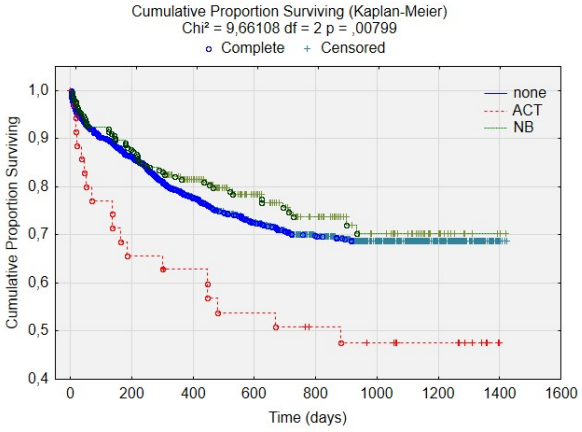


FIGURE 5.27: *Kaplan-Meier survival curve for the community programmes.*

5.2.3 Investigating predictors of readmission

From the logistic regression analysis, indicators of readmissions along with the odds ratio for the variable classes were determined. The predictive capability of the model is however inadequate (refer to Table 5.9) and is expected to perform poorly on new data. The discriminant analysis determined variables that are significant for predicting the chance of readmission, with the predictive capability of the model also being average (refer to Table 5.10), but better compared to the logistic regression model.

Although logistic regression and Cox-proportional hazards seem to provide similar results, they are fundamentally different. Cox regression displays the hazard ratio, which is the number of new cases per population-at-risk in a unit of time. The hazard function is the likeliness of a person experiencing 'death' (in this case, readmission) during the next instant, owing to having survived up to time t. Logistic regression on the other hand, evaluates the proportion of new cases that developed in a time period (thus, the cumulative incidence) and estimates the odds ratio (Pennsylvania State University, 2016).

Similar to the logistic regression and discriminant analysis results, survival time analysis also found that SA_Bi_MDD&Anx, the ACT programme and Area 4 resulted in a significantly faster readmission time. Patients with SIPD were also found to be readmitted less, and after a longer period of time when compared to patients not from this group.

CART and random forest methods are investigated further and expected to have improved classification ability. The substance dataset is not investigated further owing to substance not emerging as significant and the dataset containing too few observations throughout the various classes. Up to this point, the ungrouped dataset has produced similar results to the grouped

dataset, especially in terms of the diagnosis and area. There also exists the possibility that significant information, more specific to certain classes, might be obtained from the dataset. Accordingly, the ungrouped dataset is included in the CART analysis with the knowledge that the grouped dataset is more statistically and methodologically sound.

5.3 Building predictive models

This section presents the results of the prediction models built respectively by CART and random forests. Both methods are from the field of data mining. The predictive strength of the logistic regression and discriminant models have been discussed in Section 5.2.1.

5.3.1 Classification and regression analysis

The grouped dataset is analysed first after which the findings of the ungrouped dataset are presented. If the grouped dataset's model has a relatively good predictive strength a recommendation can be made with regard to further implementing it as a decision support tool. This is however further discussed in Chapter 6.

5.3.1.1 Grouped dataset

The CART software was set up as previously explained in Section 4.3.2 and 52 trees were constructed according to the cost sequence displayed in Figure 5.28. The figure is enlarged to include only the last few trees owing to them having the lowest CV cost and number of terminal nodes. Trees 50 and 51 were identified as optimal based on their low CV cost and low number of terminal nodes as displayed in Figure 5.28. Tree 51 was also selected by the Statistical software.

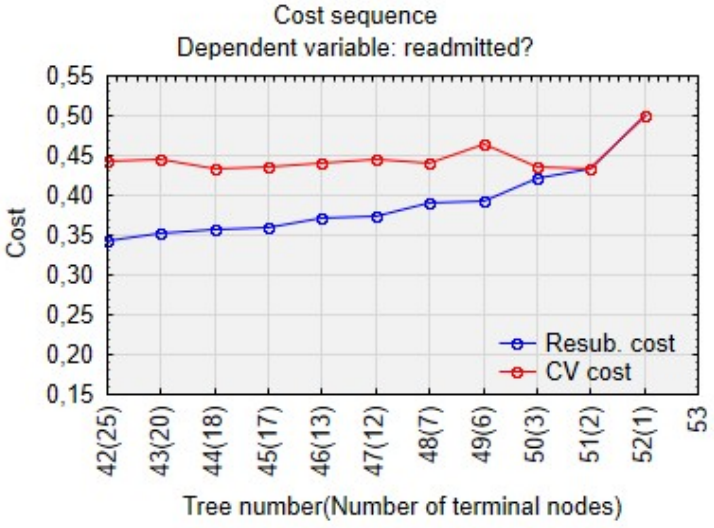


FIGURE 5.28: *Cost sequence diagram for the CART model of the grouped dataset (zoomed in).*

A summary of the predictive capability of the trees is given in Figure 5.12 and the complete results can be seen in Section G.1 of Appendix G. The trees are very similar with regard to their predictive ability, with Tree 50 providing a slightly better average at basically the same CV cost as Tree 51.

TABLE 5.12: *Summary of the predictive capability of Tree 50 and 51 respectively based on v-fold cross validation.*

	Tree 50	Tree 51
% classified correct		
1	85.47%	87.42%
0	30.11%	25.88%
average	57.8%	56.7%
Other information		
# terminal nodes	3	2
CV cost	0.435	0.434

Tree 50 conveys three splitting rules, which can be interpreted graphically from Figure G.2 in Appendix G.1. It predicts that:

1. A patient from the SA_Bi_MDD&Anx group **or** schizophrenia group **and** being younger than 19,5 years of age has a 16% of being readmitted. If chosen randomly from the dataset the readmission chance is 29%;

2. A patient from the SA_Bi_MDD&Anx group **and** being older than 19,5 years of age has a 33% chance of being readmitted, which is more than if randomly chosen from the dataset; and

3. A patient in the SIPD **or** GMC_Other group has a 16% chance of readmission.

Tree 51 has two splitting rules, displayed graphically in Figure G.1, with the second being the same as Rule 3 of Tree 50, and the first rule being similar to Rule 1 of Tree 50, but not including age in predictive rules:

1. A patient in the SA_Bi_MDD&Anx group **or** schizophrenia class is predicted to have 32% chance of being readmitted; and

2. A patient diagnosed with *SIPD* **or** belonging to the GMC_Other diagnosis group is 16% likely to be readmitted. Selecting a patient from either these two diagnoses groups thus results in a less likely chance of readmission compared to selecting a patient 'at random'.

Tree 50 ranked the variables according to importance for the model as: diagnosis, length of stay, age, area, NB/ACT and lastly follow-up which is displayed in Figure G.3 of Appendix G.1. Tree 51 ranked the variables similarly in the case of the first two variables, after which, area, NB/ACT, age and follow-up followed respectively as last, shown in Figure G.1. The area, ACT/NB and age are ranked with similar weights.

5.3.1.2 Ungrouped dataset

The CART software was set up as previously explained in Section 4.3.2 and 38 trees were constructed according to the cost sequence displayed in Appendix G.2, Figure G.5. Tree 37 was chosen as the best tree by the software which is agreeable owing to having a low CV cost (0.43) and the fewest number of terminal nodes (two). Tree 36, which has the second least terminal nodes (six), has a higher CV cost (0.47) and Tree 35, although having a lower CV cost (0.46) than Tree 36, has eight terminal nodes which gives cause to believe that in this case over-fitting occurred.

Tree 37 ranked the variables according to importance as follows: ICD10 diagnosis, LOS, follow-up, area, NB/ACT and lastly age. This is displayed in Figure G.6 of Appendix G.2. The categorised histogram displayed in Figure G.7 of Appendix G.2 that group.

In this case, Tree 37 has two splitting conditions or rules:

1. A patient diagnosed with 'GMC or other' **or** SIPD has a 16% chance of readmission, which is a better chance than that of a patient selected at random. In this case there is a 29% chance of readmission; and
2. A patient diagnosed with either schizophrenia **or** schizo-affective **or** bipolar **or** 'MDD and anxiety' has a 32% chance of being readmitted which is more than that of a patient who is selected randomly from the dataset whose readmission probability is 29%.

The predictive capability of Tree 37 is on average 56.7% based on classifying 87.4% of the readmissions and 26% of the non-readmissions correctly. Although it is regarded as more important to classify a readmission correctly as opposed to a non-readmission, 26% seems less desired.

5.3.2 Random forests

Little descriptive information can be obtained from random forest models as they are primarily used for prediction. Additionally they will be used to determine whether there is value in building a decision making tool to calculate a patient's chance of readmission. Decision rules cannot be interpreted, as with CART, owing to the model building one hundred trees and taking the majority prediction outcome of all these trees.

As previously discussed in Section 4.3.2, most settings are predetermined by formulas implemented by the software, but some settings can be experimented with to improve or cater for a solution specific to this problem. Accordingly, five tests displayed in Table 5.13 were initially conducted with each test only changing one parameter to investigate the effect it had on the classification ability. For this research it is more important to classify a patient that will be readmitted correctly (observed 1 classified as 1), than to incorrectly classify a patient that will not be readmitted as readmitted (observed 0 classified as 1). To ensure this, a higher misclassification cost can be assigned to classifying a '1 as a 0' and is tested with Run5.

A baseline test was run first with settings mostly predetermined by Statistica with a few adjustments suggested by the statistical SME as previously explained in Section 4.3.2. Thus, for the baseline test, only the prior probabilities, sub-sample proportion and test data proportion are

changed from the predetermined settings with the alternative setting for the prior probabilities being tested in Run3. Run2 experimented with the *number n cases*, Run4 had a higher *test data proportion* and Run5 a higher misclassification cost assigned to misclassifying a readmission.

TABLE 5.13: *Initial settings for each run to build a random forest model.*

Settings	Initial settings for random forest models				
	Run1	Run2	Run3	Run4	Run5
Test data proportion	0.1	0.1	0.1	0.3	0.1
Minimum n cases	39	20	39	39	39
Prior probabilities	equal	equal	estimated	equal	equal
Misclassification cost	0 1 0 1 1 1	0 1 0 1 1 1	0 1 0 1 1 1	0 1 0 1 1 1	0 1 0 1.3 1 1

TABLE 5.14: *The classification ability of the initial models.*

	Initial random forest models' results					
	% correct	Run1	Run2	Run3	Run4	Run5
Training set	0	66.89%	65.23%	99.90%	68.43%	46.88%
	1	73.32%	75.96%	2.88%	75.53%	87.02%
	average	70.11%	70.60%	51.39%	71.98%	66.95%
Test set	0	60.71%	58.93%	100.00%	62.53%	39.29%
	1	62.22%	62.22%	2.22%	53.08%	75.56%
	average	**61.47%**	**60.58%**	**51.11%**	**57.81%**	**57.43%**

The results of the first five tests is displayed in Table 5.14 and the models are evaluated based on the percentages of cases classified correctly in the training set and test set, with more consideration given to the 1's classified as 1's (readmissions). On average there is not much difference between Run1 and Run2, although Run2 classifies the 1's more accurately in the training data and thus another test, Run2B is conducted with the *minimum n cases* equal to *30* which is half way between 20 and 39. Run3, which modelled the random forest with *estimated* prior probabilities, did not provide good classification results especially with the readmissions and thus the 'equal' setting is selected. Run4 which used 30% of the data for testing performed well even though there were fewer observations to learn from (learning: 1104 data points and test: 493 data points). It is however statistically more sound to use most of the data for learning purposes. Another test will be done at 20% (Run4B) to see the effect. The test where a higher cost is associated with classifying readmission incorrectly produced the best classification ability with regard to classifying readmissions. Classifying non-readmission correctly was on average 40% and accordingly a misclassification cost of 1.2 (Run5B) will be investigated in the second round of tests displayed in Table 5.15.

TABLE 5.15: *Settings for the second round of random forest models.*

Settings	Second round's settings		
	Run2B	Run4B	Run5B
Test data proportion	0.1	0.2	0.1
Minimum n cases	30	39	39
Prior probabilities	equal	equal	equal
Misclassification cost	0 1 0 1 1 1	0 1 0 1 1 1	0 1 0 1.2 1 1

TABLE 5.16: *Predictive ability of the models built for the second round of experiments.*

	Second round's results			
	% correct	Run2B	Run4B	Run5B
Training set	0	72.85%	66.30%	51.86%
	1	71.63%	77.81%	83.17%
	average	**72.24%**	**72.06%**	**67.52%**
Test set	0	64.29%	60.26%	44.64%
	1	48.89%	58.62%	75.56%
	average	**56.59%**	**59.44%**	**60.10%**

From the results displayed in Table 5.16, Run2B indicates that with a *minimum n cases* of 30, non-readmissions are classified better and readmissions much worse. The parameter will thus be set to Statistica's recommended value, which in this case is 39 and has a higher correct classification rate of readmissions. Run4B, with a test sample of 0.2, shows improved classification with both the test and learning set compared to a test set of 0.3. Lastly, with Run5B the classification of the non-readmissions have improved to 52% in the training set, but is still low in the test set. The classification of the readmissions is still good at above 80%, where the other models achieve at best around 75%. Model 4B thus seems to be the best model, with one more experiment conducted by changing the misclassification cost of readmission to 1.1 and experimenting a last time with the test set proportion. The settings for the final round of tests are displayed in Table

5.17 with the results displayed in Table 5.18. The final tests did yield improved results with regard to being more balanced. For example test *RunFinal0.1* classified the non-readmissions better whilst still having a high classification ability for the readmissions. With each test, a trade-off is made with either classifying readmissions or non-readmissions better and with the last tests there are fewer differences on average. The final settings selected for random forest is that of *RunFinal0.1*. More data was used to build the model than to test it and a slightly higher misclassification cost was allocated to readmissions.

TABLE 5.17: *Settings for the final round of random forest models.*

Settings	Final settings	
	FunFinal0.2	RunFinal0.1
Test data proportion	0.2	0.1
Minimum n cases	39	39
Prior probabilities	equal	equal
Misclassification cost	0: 0, 1: 1.1; 1: 1, 1:	0: 0, 1: 1.1; 1: 1, 1:

TABLE 5.18: *Results for the final round of random forest models.*

	Final grouped results		
	% correct	RunFinal0.2	RunFinal0.1
Training set	0	62.64%	60.55%
	1	77.01%	79.57%
	average	**69.83%**	**70.06%**
Test set	0	55.13%	52.68%
	1	65.52%	66.67%
	average	**60.33%**	**59.68%**

For the ungrouped dataset the same settings were used as with the last model for the grouped dataset, experimenting between a 0.1 and 0.2 test data proportion and equal misclassification cost or 1.1 of readmissions. The results are displayed in Table 5.19. It was found that the model with the best predictive ability is RunU1, which has the same settings as that with which the best model with the grouped data was obtained.

TABLE 5.19: *Random Forest classification results for the ungrouped dataset.*

	Ungrouped dataset models'results				Best Grouped
	% correct	RunU1	RunU2	RunU3	
		Test set: 0.2 Misclass cost: 1.1 (readmi.)	Test set: 0.1	Test set: 0.2 Misclass cost: equal	Settings of *RunFinal0.1*
Training set	0	66.15%	63.15%	69.68%	60.55%
	1	76.96%	79.38%	72.63%	79.57%
	average	**71.56%**	**71.27%**	**71.16%**	**70.06%**
Test set	0	57.21%	53.10%	58.52%	52.68%
	1	59.78%	54.55%	54.35%	66.67%
	average	**58.50%**	**53.83%**	**56.44%**	**59.68%**

5.3.3 Comparing the predictive models

The predictive models used to analyse the data were evaluated on their classification ability. The best models for each method were chosen and are summarised in Table 5.20. For random forest the model which uses only 10% for test data is used as opposed to the other models which use all the data points to build the model.

From the results it can be seen that random forests provide the best classification results owing to classifying on average both the readmission and non-readmission better. CART and random forests have a good prediction rate with regard to readmission, which is more important in this research as explained in Section 5.3.2. The discriminant analysis results are surprisingly balanced owing to logistic regression, which is a similar method, classifying 3% of the readmissions correct.

TABLE 5.20: *Classification ability of the predictive models used in the project.*

Grouped predictive results				
% correct	Logit	Discriminant *Priors: Equal*	CART *Tree50*	Random forests *RunFinal0.1*
0	99.03%	57.83%	30.11%	60.55%
1	2.60%	58.13%	85.47%	79.57%
average	**50.8%**	**58.0%**	**57.8%**	**70.1%**

The models of the ungrouped dataset were also compared for interest sake and this comparison is displayed in Table 5.21. It is similar to the results of the grouped dataset with random forest also displaying the best results although a better indication of the model's predictive ability will be attained when the model is tested on new data. The ungrouped data has many variable classes with only a few data points as previously mentioned. Thus certain predictions may be based on this study period, for example the fourteen observations representing the patients who followed up at Tygerberg.

TABLE 5.21: *Classification ability of the predictive models from the ungrouped dataset.*

Ungrouped predictive results				
% correct	Logit	Discriminant *Priors: Equal*	CART *Tree37*	Random forests *TestU2*
0	98.24%	51.50%	25.88%	63.15%
1	6.94%	67.25%	87.42%	79.38%
average	**52.6%**	**59.4%**	**56.7%**	**71.3%**

5.4 Conclusion: Results

This chapter conveyed the implementation and results of the various statistical and data mining methodologies discussed in the previous chapter. The analyses were conducted on both the grouped and ungrouped dataset. The ungrouped dataset performed surprisingly well, but is not as statistically and methodologically sound as the grouped dataset owing to the data being skewed and unequally distributed. As mentioned, if more data could be obtained for the classes with less data, the same methods could be used to analyse the dataset.

The data was first described using basic statistics, whereafter more focus was given to determining if and to what extent the patients readmitted differed from those not readmitted. Afterwards, predictive models were built and evaluated with regard to using it for predictive modelling.

The next chapter serves the purpose of integrating the statistical results with the real-world problem and discussing the implications this may have on the decision making procedures.

CHAPTER 6

Conclusion: Results meet the real-world problem

Chapter 5 presented the results of the descriptive statistics as well as the predictive models. The data were described with various graphs and basic statistics as well as modelled in Statistica to determine significant variables pertaining to readmission. Predictive models were also built and evaluated.

This chapter serves to answer the question "so what now" by bringing the statistical results back to meet the real-world problem. This will be achieved by presenting an overview of the statistical findings and discussing the feasibility of a decision support tool before giving a summary of the project, the contributions and suggestions for further research.

6.1 Research findings and the decision-making implications

The project largely investigated variables to determine whether they are statistically valid predictors of whether patients will require readmission. The significant statistical findings that can be taken in consideration by the psychiatric clinicians when deciding which patients to discharge are summarised in this section. The feasibility of crisis-discharge decision-support based on the predictive model that was developed in this research is discussed. Finally, this study's limitations (mostly associated with the nature of the data) is discussed.

6.1.1 Overview and implications of the statistical findings

The first fundamental finding of this research, is that specific variables exist which indicate that a patient has an increased risk of requiring readmission after being crisis-discharged from the acute male ward at Stikland. The details on the variables that have been identified as predictors are summarised here.

The data analysis findings for the methods used are briefly summarised in Table 6.1 which indicates the variables and classes that play a significant role in predicting readmission at Stikland Psychiatric Hospital. The discriminant analysis results found the same significant variables as the logistic regression model and these are thus not explicitly mentioned in the table. The results

focus on the grouped dataset. The CART column indicates the variables which formed part of a splitting condition.

TABLE 6.1: *Significant variables and classes from the various analyses.*

Class	Logistic regression		CART	Survival analysis
	Significant overall?*	Significant classes		
Age			*(><19.5 years)*	
LOS				
Area	✓	Area 4[i]		
Follow up				
ACT/NB	✓	NB[d] ACT[i]		ACT
ICD10	✓	Schizophrenia[i] SA_Bi_MDD&Anx[i] SIPD[d]	Schizophrenia SA_Bi_MDD&Anx SIPD GMC and other	SA_Bi_MDD&Anx
Substance		Tik[i]		

[i]increased probability for readmission [d]decreased probability for readmission
*Discriminant analysis determined the same variables to be significant

From the dataset pertaining to mainly the first admissions of patients admitted during 2012 to 2014 a number of statistical significant trends were found. Patient diagnosis emerged as significant in all the analyses conducted, especially the SA_Bi_MDD&Anxiety class. From the results of the ungrouped dataset the most contributing factors are schizo-affective and bipolar disease which leads to patients being three times more likely to be readmitted than patients not suffering from either disorder. Schizophrenic patients are also twice as likely to be readmitted as patients who do not have the disorder. Similarly, the survival analysis found that the SA_Bi_MDD&Anx group followed by schizophrenia had more readmissions in a shorter period of time as compared to the other groups.

The CART tree built on the grouped dataset split on the diagnosis and patient age, and predicts that fewer patients are readmitted if they are from the SA_Bi_MDD&Anx or schizophrenia group and are younger than 19.5 years (16% probability of readmission compared to 29% if selected randomly from dataset). The same diagnosis group and a patient older than 19.5 years showed an increased chance for readmission (33% probability of readmission). The third rule indicated that a patient with SIPD or from the 'GMC and other' group had a lower chance for readmission (16% probability of readmission).

The community program variable was also found to be significant at 10% in the logistic regression analysis which indicated that patients from New Beginnings are half as likely to be readmitted as patients not from the group. The ACT-patients were found to more than twice as likely to be readmitted when compared to non-ACT patients, which is alarming, but as explained previously, the programme focusses on acute patients and aims to reduce the LOS and number of readmissions. By analysing only the one admission and not following patients individually may inaccurately appear to contribute negatively to readmission. ACT patients also seem to be readmitted much faster than patients not from the class when comparing the Kaplan-Meier graphs for the variable.

Of the area-variable, Area 4, which represents the direct admissions was found to be significant, with patients from this class being 1.7 times more likely of readmission than patients not from the group. As mentioned, this cannot be seen as an indication of the quality of care provided at Stikland Psychiatric Hospital, owing to patients from the Stikland area having direct access to the hospital, whereas other patients first need to go to their primary health care clinic or doctor for observation and only then be referred to Stikland Psychiatric Hospital. Patients from Area 4 also show more readmissions in a much shorter time compared to other areas when analysing the Kaplan-Meier survival curves.

6.1.2 Summary of the predictive results

The second fundamental finding of this research, is that the variables that have a significant influence on risk of readmission following crisis-discharge, can be harnessed to create predictive models with promising levels of accuracy.

A random forest model was constructed and provides the best prediction results compared to the other methods used in this research (CART, discriminant analysis and logistic regression). On average the model predicted 70% of the observations correctly (classifying 79.6% of the readmission and 61.6% of the non-readmission correctly). The model was constructed from data of only three years, using 10% as a test set. Of the observations in the test set 66.7% of the readmissions, and 52% of the non-readmissions were classified correctly. The predictive results suggest that it may prove valuable to further this research.

6.1.3 Feasibility of a decision support tool

The prediction models definitely show potential with regard to predicting a patient's chance for readmission. Although the predictive models are evaluated on the number of correctly classified cases, each classification stems from a probability which is rounded up or down for a readmission and non-readmission respectively. With the decision support tool the probability will be used as output, and not the classification. This will enable the clinicians to compare patients and select the patient with the lowest probability of readmission to be discharged. The random forest model had the best prediction results, which was expected by the statistical SME, and accordingly additional Statistica software can be downloaded to capture the code generated in the background and used to develop a user interface with programming software such as C++. A programmer will have to be appointed to modify the code specific to the programming software, build an interface for the clinical user to input data as well as generate results with regard to a patient's probability of readmission based on the prediction model's calculations.

The detail pertaining to developing and building the decision support tool is not discussed in this research. It is envisaged that the clinical SME will input patient variables that are used in the prediction model (diagnosis, age, follow-up, LOS to date, area and participation in community programmes) and that the decision tool will then output the patient's probability of readmission. It might also be useful to input numerous patients' information and then simultaneously provide the results and rank the patients according to the probability for readmission. The details pertaining to modifying the tool for the user will have to be researched along with the various other aspects pertaining to developing a decision support tool.

The decision tool will most probably require the psychiatrists to enter the data manually into the tool. On average, about three to five patients have to be selected daily to be discharged. The psychiatrist will decide either to input all patients that could be discharged or only a select few they may think best to choose from. The possibility of extracting data directly from Clinicom into the tool will have to be investigated. It is however expected to be infeasible owing to the information in the worksheets being inconsistent and in some cases incomplete; the information not being grouped in the same classes used in the predictive model; and information such as community programmes not being contained in the Excel sheets (which is inferred from the 2012-2014 workbooks). The data capturing process may have improved since 2014, but it still seems unlikely that that it would be feasible to directly link the two software programs. The decision tool may allow the input of '.csv' files, which can be generated from the Excel sheets sourced from Clincom, but these sheets will also first be cleaned and reviewed. The clinical SMEs however believe that it is feasible and that implementing and building a decision support tool should be investigated and considered in the near future.

6.1.4 Data limitations

The dataset did have some limitations, but this is not uncommon for medical data. The various classes in the variables, which were initially grouped by the clinical SMEs, indicated the need for further grouping owing to classes still being unequally divided with some classes containing, for example, as little as 1% of the total observations and another group containing 66%. This influences the statistical validity of the results owing to them pertaining to a very small sample which may be used to make discharge decisions in the future.

It is believed that with more data, other smaller variable groups may emerge to present significant information and contribute to prediction models and the variables already found significant will emerge again to be significant. It is also suspected that with more observations about the substance use more conclusive results could be obtained.

There is evidence that suggests the data capturing process changed slightly from 2012 to 2014, which suggests the possibility of inconsistent data. This resulted in the inability to include some variables which may have been interesting to analyse, for example marital status and income. The trends between patients' admissions could not be analysed owing to having too few data points. It is however expected that to analyse these finer trends between (re)admissions data from a wider scope than Stikland Psychiatric Hospital will have to be included.

Owing to having a relatively small dataset for building prediction models, a test set could not be extracted from the dataset and used to calculate a model's prediction accuracy owing to it being better to build a prediction model from the most data possibly available. This however did not restrict the development of prediction models and it was still possible to evaluate the prediction ability. It might be useful to test the prediction models on their ability to classify 2015's admission data before committing to developing a decision support model.

6.2 Contributions of this research

This project contributes to both the psychiatric health care sector, especially Stikland Psychiatric Hospital, and to academic literature:

1. It contributes to academic literature by analysing admission and discharge data from a psychiatric hospital with statistical and data mining methods, with CART and random forest being novel data analysis approaches (to our knowledge);
2. It documents the general operational and decision policies currently implemented at Stikland Psychiatric Hospital and describes the psychiatric sector of South Africa;
3. It evaluates the feasibility of the data being used to build a decision support model and finds it to be viable, especially with the random forest model, which is a novel approach (to our knowledge);
4. The project contributes novel information with regard to readmission trends and indicators specifically from the population served by Stikland Psychiatric Hospital, confirming clinical SME's expectations as well as providing new insights in others; and
5. The limitations found during this project with regard to analysing data specifically from a psychiatric health institution were presented and may inform future research.

6.3 Opportunities for further work

The following opportunities potentially may add value to crisis-discharge decisions and the general understanding of factors that significantly contribute to readmission:

1. Including more data, for example from 2015 and re-running the various analysis methods to determine whether the smaller groups may emerge as significant. That is if the new data added more observations to the groups that are currently too small to make statistically valid conclusion from;
2. Using 2015 or new data as a test set to evaluate the predictive ability of the prediction models, especially random forests and CART. The ability of the model to accurately predict new data improves with more observations, which can be seen in Table 6.2 (new data is represented by a test set which comprises 10% of the data set);

TABLE 6.2: *Predictive capability of models with varying sample sizes.*

	Predictive capability						
	% correct	2012		2012-2013		2012-2014	
Test set	0	63.04%	n = 642*	50.75%	n = 1176*	52.68%	n = 1598*
	1	46.15%		61.02%		66.67%	
	average	**54.56%**		**55.885%**		**59.68%**	

*Total size of the data set

3. Gathering data about the substance abuse patients which is currently incomplete, including it in the grouped-dataset and rerunning the analyses in order to obtain more valid results with regard to substance abuse and readmission; and

4. Formally evaluating the feasibility of practically implementing a decision support tool. If found valuable and feasible from the user's point of view, the random forest prediction model's code could be used and implemented in a decision support tool for daily use by the psychiatric clinicians.

6.4 Project summary

This project focused on analysing readmission at Stikland Psychiatric Hospital in order to determine indicators for readmission and possibly improve decision making with regard to selecting patients for discharge who have the least probability of readmission. The feasibility of a predictive model and decision support tool was also evaluated.

Chapter 1 introduced the project, rationale and the research design and plan, whereafter Chapter 2 and Chapter 3 presented background to the research. Chapter 2 introduced the general context of the project by briefly describing the South African healthcare sector as well as psychiatric care in South Africa, specifically the Western Cape and Stikland Psychiatric Hospital. Readmission and deinstitutionalisation were also introduced in the chapter along with published studies that similarly, as in this project, investigated indicators for readmission within certain populations.

The concept 'learning from data' was introduced in Chapter 3 pertaining mainly to field of data mining, which also entails statistical measures implemented in the study. For substantiation, similar studies published with regard to the data analysis techniques employed in the study were presented. Focus was given to analysis techniques most applicable to this project from analysing these studies, literature and a statistical SME's input.

In Chapter 4 the chosen techniques were further explained from a more practical point of view, specifically with regard to applying them to this project's dataset. The data management and cleaning process were also presented along with details pertaining to the methods used as well as the rationale for the variables included, excluded or grouped for this project. The variables included in the similar published studies were also presented. The methods were applied to the data and the results pertaining to readmission at the acute male ward were presented in Chapter 5. Chapter 6 discussed the project's contributions, the possibility of a decision support tool and future work.

6.5 Conclusion

This chapter served as the closing chapter of this project and provided a brief summary of the results, the practical implications and limitations, and the feasibility of developing a decision support tool. The contributions of this research, opportunities for further research and finally an overview of the content of the project are also presented in this chapter.

References

Agarwal, S. & Tomar, D. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, **5**, 241–266.

Agency for Healthcare Research and Quality (2014). Management strategies to reduce psychiatric readmissions. Available at: `www.effectivehealthcare.ahrq.gov`.

Aljumah, A.A., Ahamad, M.G. & Siddiqui, M.K. (2012). Application of data mining: Diabetes healthcare in young and old patients. *Journal of King Saud University - Computer and Information Sciences*, **25**, 127–136.

Barekatain, M., Maracy, M., Hassannejad, R. & Hosseini, R. (2013). Factors associated with readmission of patients at a university hospital psychiatric ward in Iran. *Psychiatry journal*, **2013**, 5.

Barron, P. & Pillay, Y. (2014). Progress towards the Millennium Development Goals in SA (foreword). *South African Medical Journal*, **104**, 223.

Bate, A., Lindquist, M., Edwards, I., Olsson, S., Orre, R., Lansner, A. & De Freitas, R.M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, **54**, 315–321.

Bateman, C. (2014). Mental health under-budgeting undermining SA's economy. *South African Medical Journal*, **105**, 7.

Bellazzi, R. & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, **77**, 81–97.

Bernardo, a.C. & Forchuk, C. (2001). Factors associated with readmission to a psychiatric facility. *Psychiatric services (Washington, D.C.)*, **52**, 1100–1102.

Botha, U., Koen, L., Oosthuizen, P., Joska, J. & Hering, L. (2008). Assertive community treatment in the South African context. *African journal of psychiatry*, **11**, 272–275.

Botha, U.A., Koen, L., Joska, J.A., Hering, L.M. & Oosthuizen, P.P. (2010). Assessing the efficacy of a modified assertive community-based treatment programme in a developing country. *BMC psychiatry*, **10**.

Bramer, M. (2007). *Principles of data mining*. Springer, London, 1st edn.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1993). *Classification and regression trees*. Chapman & Hall, New York, 2nd edn.

Byrne, S.L., Hooke, G.R. & Page, A.C. (2010). Readmission: A useful indicator of the quality of inpatient psychiatric care. *Journal of Affective Disorders*, **126**, 206–213.

Chiumia, S. & van Wyk, A. (2014). Do a third of South Africans really suffer from mental ilnesses. Available at: `https://africacheck.org/reports/do-a-third-of-south-africans-really-suffer-from-mental-illnesses/`.

Chopra, M., Lawn, J.E., Sanders, D., Barron, P., Karim, S.S.A., Bradshaw, D., Jewkes, R., Karim, Q.A., Flisher, A.J., Mayosi, B.M., Tollman, S.M., Churchyard, G.J. & Coovadia, H. (2009). Achieving the health Millennium Development Goals for South Africa: challenges and priorities. *The Lancet*, **374**, 1023–1031.

Cullinan, K. (2006). Health services in South Africa: a basic introduction. Tech. Rep. Health-e News Service.

Curiac, D.I., Vasile, G., Banias, O., Volosencu, C. & Albu, A. (2009). Bayesian network model for diagnosis of psychiatric diseases. *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces*, 61–66.

Dell (2015a). Classification and regression trees. Tech. rep., Dell.

Dell (2015b). Statistica Textbook: Discriminant Function Analysis. Available at: `http://documents.software.dell.com/Statistics/Textbook/Discriminant-Function-Analysis`.

Department of Health (2013). National Mental Health Policy Framework and Strategic Plan 2013-2020. Tech. rep.

Department of Health (2014). Strategic plan 2014/15 - 2018/19. Tech. rep., Pretoria.

Durbin, J., Lin, E., Layne, C. & Teed, M. (2007). Is readmission a valid indicator of the quality of inpatient psychiatric care? *Journal of Behavioral Health Services and Research*, **34**, 137–150.

Econex (2013). The South African private healthcare sector: Role and contribution to the economy. Tech. rep.

Elliott, A. & Woodward, W. (2007). *Statistical analysis quick reference guidebook*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America.

Field, A. (2009). *Discovering statistics using SPSS*, vol. 58. SAGE Publications Ltd, 2nd edn.

Gillis, L.S., Sandler, R., Jakoet, A. & Elk, R. (1986). Readmissions to a psychiatric hospital. *South African medical journal*, **70**, 735–739.

Goldberg, J. (2014). Mental health: Types of mental illness. Available at: `http://www.webmd.com/mental-health/mental-health-types-illness`.

Han, J. & Kamber, M. (2006). *Data mining concepts and techniques*. Elsevier, San Francisco, 2nd edn.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning*. Springer, USA, 2nd edn.

Haywood, T.W., Kravitz, H.M., Grossman, L.S., Cavanaugh, J.L., Jr JM, D. & Lewis, D.a. (1995). Predicting the "revoving door" phenomenon among patients with schizophrenic, schizoaffective, and affective disorders. *American Journal of Psychiatry*, **152**, 856–861.

Health Systems Trust (2012). The National Health Care Facilities Baseline Audit. Tech. rep.

Health Systems Trust (2014). South African Health Review 2013/14. Tech. rep., Health Systems Trust, Durban.

Heggestad, T. (2001). Operating conditions of psychiatric hospitals and early readmission–effects of high patient turnover. *Acta psychiatrica Scandinavica*, **103**, 196–202.

Herman, A.a., Stein, D.J., Seedat, S., Heeringa, S.G., Moomal, H. & Williams, D.R. (2009). The South African Stress and Health study: 12-month and lifetime prevalence of common mental disorders. *South African Medical Journal*, **99**, 339–344.

Heslin, K.C., Ph, D., Weiss, A.J. & Ph, D. (2015). Hospital readmissions involving psychiatric disorders. Tech. Rep. 13, Agency for Healthcare Research and Quality.

Human, A. (2010). A tale of two tiers: Inequality in South Africa's healthcare system. *University of British Columbia Medical Journal*, **2**, 33.

Innes, H., Lewsey, J. & Smith, D.J. (2015). Predictors of admission and readmission to hospital for major depression: A community cohort study of 52,990 individuals. *Journal of Affective Disorders*, **183**, 10–14.

Izad Shenas, S.A., Raahemi, B., Hossein Tekieh, M. & Kuziemsky, C. (2014). Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes. *Computers in Biology and Medicine*, **53**, 9–18.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques*. Springer, USA, 1st edn.

Jacob, S.G. & Ramani, R.G. (2012). Data mining in clinical data sets: A review. *International Journal of Applied Information Systems*, **4**, 15–26.

Janse Van Rensburg, A.B.R. (2007). A framework for current public mental health care practice in South Africa. *African journal of psychiatry*, **10**, 205–209.

Johnstone, P. & Zolese, G. (1999). Length of hospitalisation for people with severe mental illness (Review). Tech. Rep. 2.

Jones, R., Yates, W.R. & Zhou, M.H. (2002). Readmission rates for adjustment disorders: Comparison with other mood disorders. *Journal of Affective Disorders*, **71**, 199–203.

KHUMALO, G. (2012). Call for mental health service rethink. Available at: http://www.southafrica.info/about/health/mental-130412.htm.

KIDD, M. (2016a). Meeting: ANOVA assumptions and logistic regression, 3 June, Stellenbosch.

KIDD, M. (2016b). Meeting: Descriptive statistics in Statistica and cleaning the dataset, 13 June, Stellenbosch.

KIDD, M. (2016c). Statisical SME meeting: Reviewing the results, 15 July, Stellenbosch.

KIDD, M. & SMIT, I.M. (2016). Meeting: Combining variables in the dataset, 6 June, Stellenbosch.

KOEN, L. (2016a). Email: LOS and days discharged, 1 August, liezlk@sun.ac.za.

KOEN, L. (2016b). Email: Readmisson within 30 and 90 days, 5 August, liezlk@sun.ac.za.

KOEN, L. & SMIT, I.M. (2016a). Meeting: Finalising the dataset, 29 March, Stikland Hospital.

KOEN, L. & SMIT, I.M. (2016b). Meeting: Results discussion, 19 July, Stikland Hospital.

KOEN, L. & SMIT, I.M. (2016c). Meeting: Reviewing initial results, 9 June, Stikland Hospital.

KOH, H.C. & TAN, G. (2005). Data mining applications in healthcare. *Journal of healthcare information management*, **19**, 64–72.

KRISHNAN, K.R.R. (2010). Data mining in healthcare. In *IndiaCom-2010 4th National conference on "computing for national development"*, Chennai.

KUHN, M. & JOHNSON, K. (2013). *Applied predictive modeling*. Springer, New York, 1st edn.

LAERD STATISTICS (2013a). Binomial logistic regression using SPSS Statistics. Available at: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php.

LAERD STATISTICS (2013b). Independent t-test for two samples. Available at: https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php.

LAERD STATISTICS (2013c). Testing for normality using SPSS statistics. Available at: https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php.

LAERD STATISTICS (2015). Mann-Whitney U Test using SPSS Statistics. Available at: https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php.

LAROSE, D.T. (2005). *Discovering knowledge in data: An introduction to data mining*. John Wiley and SOns Inc., New Jersey, 1st edn.

LEMKE, K. (2012). Building a predictive model for 30-day inpatient readmission using PROC PHREG. 1–13.

Li, R. (2015). Top 10 data mining algorithms. *KD Nuggets*.

Loch, A.A. (2012). Stigma and higher rates of psychiatric re-hospitalization: São Paulo public mental health system. *Revista Brasileira de Psiquiatria*, **34**, 185–192.

Lyons, J.S., O'Mahoney, M.T., Miller, S.I., Neme, J., Kabat, J. & Miller, F. (1997). Predicting readmission to the psychiatric hospital in a managed care environment: implications for quality indicators. *The American journal of psychiatry*, **154**, 337–40.

Malesu, R.R. (n.d.). Factors influencing readmissions of schizophrenics to psychiatric hospital.

Mark, T.L., Vandivort-Warren, R. & Montejano, L.B. (2006). Factors affecting detoxification readmission: Analysis of public sector data from three states. *Journal of Substance Abuse Treatment*, **31**, 439–445.

Marques de Sá, J. (2007). *Applied Statistics using SPSS, Statistica, Matlab and R*. Springer, New York, 2nd edn.

Mayo Clinic (2014). Mental illness. Available at: `http://www.mayoclinic.org/diseases-conditions/mental-illness/basics/definition/con-20033813?DSECTION=all{\&}p=1`.

Mayoral, F., Perez, O., Romero, M., Hernandez, J. & Rius, F. (2012). Poster #261 A prospective study of factors associated to readmissions in adult schizophrenic patients discharged from a short-stay psychiatric inpatient unit of a university general hospital.

Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q. & Liu, Q. (2012). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung Journal of Medical Sciences*, **29**, 93–99.

Moss, J., Li, A., Tobin, J., Weinstein, I.S., Harimoto, T. & Lanctôt, K.L. (2014). Predictors of readmission to a psychiatry inpatient unit. *Comprehensive Psychiatry*, **55**, 426–430.

Motsoaledi, A. (2012). Progress and changes in the South African health sector. *The Lancet*, **380**, 1969–1970.

News24 (2013). 84 % of South Africans get 2nd rate healthcare - Motsoaledi. Available at: `http://www.news24.com/SouthAfrica/News/84-of-South-Africans-get-2nd-rate-healthcare-Motsoaledi-20130912`.

Ngoepe, K. (2016). SA has one of the world's most expensive private healthcare systems WHO. *News24.com*.

Niehaus, D.J.H., Koen, L., Galal, U., Dhansay, K., Oosthuizen, P.P., Emsley, R.A. & Jordaan, E. (2008). Crisis discharges and readmission risk in acute psychiatric male inpatients. *BMC psychiatry*, **8**, 44.

Paramasivam, V., Yee, T.S., Dhillon, S.K. & Sidhu, A.S. (2014). A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. *Biocybernetics and Biomedical Engineering*, **34**, 139–145.

Patel, B. (2014). Message from the director. *SA Federation for Mental Health*.

Pennsylvania State University (2016). Lesson 13: Proportional Hazards Regression (STAT 507). Available at: `https://onlinecourses.science.psu.edu/stat507/node/81`.

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Available at: `http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html`.

Powell, T.M. & Bagnell, M.E. (2012). Your survival guide to using time-dependent covariates. Tech. rep.

Ramlall, S. (2012). The Mental Health Care Act No 17 - South Africa. Trials and triumphs: 2002-2012. *African journal of psychiatry*, **15**, 407–410.

Riffenburgh, R.H. (2012). *Statistics in medicine*. Elsevier Inc., USA, 3rd edn.

Ruff, B., Mzimba, M., Hendrie, S. & Broomberg, J. (2011). Reflections on health-care reforms in South Africa. *Journal of public health policy*, **32 Suppl 1**, S184–S192.

Smit, I.M. (2016). Interview: Community care programmes and Stikland Hospital, 29 March, ingesmit@sun.ac.za.

Sorsdahl, K., Stein, D.J. & Lund, C. (2012). Mental health services in South Africa: scaling up and future directions. *African journal of psychiatry*, **15**, 168–71.

South African College of Applied Psychology (2013). Mental health in south africa: whose problem is it? Available at: `http://www.sacap.edu.za/blog/uncategorized/mental-health-south-africa-whose-problem-counselling/`.

South African Federation of Mental Health (2011). Mental Illness Info Pack July 2011. Available at: `http://www.safmh.org.za/Images/understandingMentalIllness.pdf`.

South African Federation of Mental Health (2013). JULY AWARENESS CAMPAIGN. Tech. rep., SAFMH.

Statistica (2015a). Classification and Regression Trees (C&RT) - Computational Details. Available at: `http://documentation.statsoft.com/STATISTICAHelp.aspx?path=GXX/Gcrt/Overviews/ComputationalDetails`.

Statistica (2015b). Models for Data Mining. Available at: `http://documentation.statsoft.com/STATISTICAHelp.aspx?path=glossary/GlossaryTwo/M/ModelsforDataMining`.

Statistica (2015c). Survival analysis: Cox regression models. Available at: `http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Survival/SurvivalAnalysis/Examples/Example4RegressionModels`.

Statsoft (2013). Stastistica formula guide: Logistic regression. Available at: `http://documentation.statsoft.com/portals/0/formulaguide/LogisticRegressionFormulaGuide.pdf`.

Statsoft (2015). Statistics textbook - Random forests. Available at: http://www.statsoft.com/Textbook/Random-Forest.

Stein, D.J. (2014). A new mental health policy for South Africa. *South African Medical Journal*, **104**, 115.

Stikland Hospital (n.d.). Stikland Hospital. Available at: http://www.stiklandhospital.co.za/.

Stroman, D.F.. (2003). *The disability rights movement: from deinstitutionalization to self-determination.*. University Press of America, 2003, University of Michigan.

Sullivan, G.M. & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education*, **4**, 279–82.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, **19**, 227–229.

Thomas, E., Cloete, K.J., Kidd, M. & Lategan, H. (2015). A decentralised model of psychiatric care: Profile, length of stay and outcome of mental healthcare users admitted to a district-level public hospital in the Western Cape. *South African Journal of Psychiatry*, **21**, 8.

UCLA: Statistical Consulting Group (2011). Introduction to SAS. Likelihood ration, Wald and Lagrange multiplier test. Available at: http://www.ats.ucla.edu/stat/mult{_}pkg/faq/general/nested{_}tests.htm.

Veloso, R., Portela, F., Santos, M.F., Silva, Á., Rua, F., Abelha, A. & Machado, J. (2014). A clustering approach for predicting readmissions in intensive medicine. *Procedia Technology*, **16**, 1307–1316.

Weich, L. & Pienaar, W. (2009). Occurrence of comorbid substance use disorders among acute psychiatric inpatients at Stikland Hospital in the Western Cape, South Africa. 213–217.

Western Cape Government (2011). 2020 The future of health care in the Western Cape. *Western Cape Government*.

WHO (2007). WHO-AIMS report on mental health system in South Africa. Tech. rep., WHO, Cape Town, South Africa.

Wickizer, T.M. & Lessler, D. (1998). Do treatment restrictions imposed by utilization management increase the likelihood of readmission for psychiatric patients? *Medical care*, **36**, 844–850.

World Health Organization (n.d.). Millennium Development Goals. Available at: http://www.who.int/topics/millennium{_}development{_}goals/about/en/.

Yussuf, A., Kuranga, S., Balogun, O., Ajiboye, P., Issa, B., Adegunloye, O. & Parakoyi, M. (2008). Predictors of psychiatric readmissions to the psychiatric unit of a tertiary health facility in a Nigerian city. A 5-year study. *African Journal of Psychiatry*, **11**, 187–190.

APPENDIX A

Additional information about the real-world problem

This appendix contains additional information on topics presented and discussed in Chapter 2.

A.1 Less common mental illnesses

Section 2.2.1 introduced the more common mental illnesses. Some of the more uncommon mental illnesses are:

- Stress response syndromes (adjustment disorders) which may occur within three months from a stressful event and end within six months;
- Dissociative disorders which involve a person suffering from extensive changes in memory, consciousness or general awareness of themselves and surroundings. It may be linked to sudden and extreme stress;
- Factitious disorders where a person creates or complains of physical and/or emotional symptoms on purpose in order to get help or attention;
- Somatic symptom disorders where a person claims to experience physical symptoms of an illness although a doctor cannot find a cause; and
- Sexual and gender disorders which affects a person's sexual behaviour and desires (Goldberg, 2014).

A.2 Healthcare facilities in South Africa

In this section, a basic description will be given firstly on the various public healthcare facilities in South Africa and, secondly, the types of psychiatric institutions in South Africa. The healthcare facilities were briefly introduced in Section 2.2.3.

A.2.1 Public healthcare facilities

The basic point of entry to health services in South Africa is at primary level by means of local clinics and community health centres where 'ambulatory patients', people who can walk and do not require a bed, are treated. Clinics are facilities with a range of primary healthcare services provided eight hours per day. Certain staff members are in some cases required to sleep at the facility in case of emergency. A Community health centre, in turn, also provides primary services along with 24-hour maternity and emergency services. The facility will have up to 30 beds and a patient may be admitted for observation for a maximum time of 48 hours. The patients will not be admitted as inpatients and the centre does not have an operating theatre or give general anaesthetics (Cullinan, 2006).

Primary healthcare services are generally managed by nurses with regular visits from doctors. Patients requiring higher levels of care have to be referred to a secondary level centre or hospital. Primary-level services are focused on providing preventative, educative (advertisements), curative and rehabilitative care such as care for mothers and children, family planning, treatment for sexually transmitted diseases, immunisation, treating minor trauma as well as providing services with regard to chronic illnesses such as diabetes or high blood pressure (Cullinan, 2006).

Hospitals are important for providing inpatient care, but also serve outpatients by means of emergency departments. South African public hospitals are generally categorised as either District, Regional or Tertiary. The government has similarly started to refer to hospitals as being level 1, 2 or 3 hospitals. District or level 1 hospitals are facilities which are open 24 hours, seven days a week with a variety of outpatient and inpatient services. These hospitals have an operating theatre, emergency services and between 30 and 200 beds. A level 1 hospital should provide diagnosis, treatment, counselling and rehabilitation services. At generalist level it should provide family medicine, primary healthcare, obstetrics, psychiatry, rehabilitation, surgery, paediatrics, geriatrics and medicine. The hospital will most probably not have an ICU and may provide only general anaesthesia (Cullinan, 2006).

Regional or level 2 hospitals provide care which requires general practitioners as well as specialists. A hospital providing only one specialist service will be classified as a 'specialised level 2 hospital'. A level 2 hospital should have staff permanently employed in at least five out of eight specialties, namely medicine, surgery, paediatrics, orthopaedics, anaesthetics, diagnostic radiology, psychiatry, obstetrics and gynaecology (Cullinan, 2006).

Level 3 hospitals, which are again categorised as either provincial tertiary (tertiary 1), national referral (tertiary 2), central referral (tertiary 3) or specialised provides specialist as well as subspecialist care where the types of services are also categorised into three groups, as seen in Table A.1 (Cullinan, 2006).

Provincial tertiary hospitals work in close conjunction with regional hospitals, providing mostly level-three care which involves the expertise of clinicians who specialise in sub-specialties of group 1, e.g. sub-specialties within surgery include neurosurgery and urology. A general hospital will provide at least half of the range of sub-specialties in group one where a specialised hospital will specialise in one or two of the specialties in groups one, two or three (Cullinan, 2006).

National Referral Hospitals are tertiary 1 hospitals which also provide a range of other specialised services - such as those from group two. Central referral hospitals, of which there are only a

few, will also have a group of sub-specialties from group three and are equipped with highly specialised units able to provide multi-specialised services, research and innovation. The quality of care will be more comprehensive, but will serve lower volumes at higher costs (Cullinan, 2006).

Specialised hospitals (level 3) focus on one specialty such as maternal care, cardiology and infectious diseases. Two common examples that focus on popular chronic diseases are psychiatric and TB hospitals that provide long-term inpatient care (Cullinan, 2006).

TABLE A.1: *Specialties of tertiary hospitals in South Africa. (Reproduced from: (Cullinan, 2006).)*

Group 1 specialties	**Group 2 specialties**	**Group 3 specialties**
Anaesthetics	Cardiology	Hepatology
Burns	Cardiothoracic surgery	Liver transplant
Clinical pharmacology	Clinical immunology	
Critical and intensive care	Craniofacial surgery	
Dermatology	Endocrinology	
Diagnostic radiology	Geriatrics	
Ear, nose and throat	Haematology	
Gastroenterology	Human genetics	
Infectious diseases	Medical and radiation oncology	
Mental health	Neurology	
Neonatology	Nuclear medicine	
Obstetrics and gynaecology	Paediatric sub-specialities	
Ophthalmology	Renal transplant	
Orthopaedics	Rheumatology	
Paediatric medicine	Spinal injuries	
Paediatric surgery and intensive care		
Plastic and reconstructive surgery		
Rehabilitation centre		
Respiratory medicine		
Trauma		
Urology		
Vascular surgery		

A.2.2 Psychiatric healthcare facilities

The psychiatric healthcare facilities in South Africa were briefly mentioned in Section 2.2.3 and are further described in this section.

Outpatient facilities: In 2007, there were 3 460 outpatient facilities in the country. Records are rarely kept of diagnoses of people treated in outpatient facilities. The gender of the outpatients and whether they were children were unknown. In rare cases where files were kept, the files were only the individual case files and not utilised for service planning. Data for two provinces provided the average number of contacts per user, which was 1.7 and 12 for the Western Cape and North West, respectively. Some provinces reported that follow-up care is provided at all their outpatient facilities while other provinces stated that no follow-up is provided. This means that an average of 44% of outpatient facilities in South Africa provide follow-up care. About 1 660 users per 100 000 population are treated annually in these facilities, calculated with data from only four provinces (WHO, 2007).

Day treatment facilities: The country has 80-day treatment facilities, of which about half are administered by the South African Federation for Mental Health (SAFMH). Up to 2007, there were no day facilities only for children and teens. The SAFMH claims that 41% of the patients at day facilities are female and there are no children or teens. The Department of Health (DoH) do not keep any statistics on gender and age in these facilities (WHO, 2007). In a report posted by the SAFMH in June 2013, the figures were still unchanged and 3.4 users per 100 000 population are treated in these facilities.

Community-based inpatient units: These are units within general hospitals. South Africa has about 41 community-based psychiatric inpatient units, but no records about the gender, age and diagnoses are kept. The average length of admission is also not recorded on a general basis. There are a total of 2.8 beds per 100 000 population. Of the beds in this group, 3.8% are reserved for children and teens (WHO, 2007).

Community residential facilities: South Africa has 63 residential-type facilities of which 47% were started by the SAFMH and 3.6 beds per 100 000 of the population are provided by these facilities. The average number of days spent at these facilities is unclear, with three organisations reporting data that resulted in an average annual length of stay of 364 days. In the SAFMH facilities of this kind, 41% of the patients are female (WHO, 2007).

Mental health hospitals: There are 23 psychiatric hospitals in South Africa, which provide 18 beds per 100 000 population. Of these hospitals, 79% of these hospitals work together with outpatient facilities. From 2002 to 2007, the amount of beds has declined on average with 7.7%, although there exists high variability between the provinces - with provinces such as the Free State having an increase in the amount of beds and other provinces such as the Western Cape (21%), Eastern Cape (27%) and Limpopo (21%) experiencing a decrease (WHO, 2007).

The DoH no not keep records of diagnoses and records of the number of users treated were only provided by the Western Cape and North West, which was 318 per 100 000 population (WHO, 2007). The average length of stay in the Western Cape is 32 days (WHO, 2007).

Forensic and other: There are about 1 676 beds in forensic inpatient units for people with mental disorders and 1 930 beds in other homes, for example for people with learning disabilities, the destitute and detoxification facilities (WHO, 2007). There are 3.5 beds per 100 000 population in forensic facilities (South African Federation of Mental Health, 2013).

A.3 Challenges and status of psychiatric hospitals in the Western Cape

Mental healthcare in the Western Cape were introduced in Section 2.2.6.1. Table A.2 summarises the challenges and status of psychiatric hospitals from what was planned for in 2010, the status quo in 2011 and then the envisaged plan for 2020 (Western Cape Government, 2011).

TABLE A.2: *Western Cape Strategic Plans for Mental Health Hospitals from 2010 to 2020. (Adapted from: (Western Cape Government, 2011).)*

Planned for 2010	Status Quo in 2011	Goals for 2020
• Implementing the 2002 Mental Healthcare Act and Mental Health Review Board. • Distinguishing between psychiatric and intellectual disability (IDS) patients. • De-hospitalising long-term IDS and psychiatric patients. • Increasing the capacity for acute patients altogether and psychiatric health services in particular. • Developing models for managing discharged patients within communities.	• Experienced pressure on capacity owing to co-morbidity linked to substance abuse and to support de-hospitalised care. • Developed innovative models in order to support chronic patients that are discharged into the community. Intellectual disability (IDS): • Patients are treated at alternative facilities rather than hospitals. • The care methods are not well defined yet in these facilities.	• Clearly defined methods of care and outcomes for each geographic service area. • Standardised methods and pathways. • Have general mental service capacity on all levels of care. • IDS residential care must be managed in the general social network than by health providers with exceptions to patients with co-morbid diagnoses or specific requirements.

APPENDIX B

More detail on the science of learning from data

This appendix presents additional information to topics that were introduced or mentioned in Chapter 3.

B.1 Data mining tasks

This section introduces the most common tasks carried out by data mining first introduced in Section 3.1.1.

Description is used for describing trends and patterns in the data. The model should provide clear patterns that correspond to intuitive interpretation. High-quality description can often be achieved by graphically exploring data (Larose, 2005).

Classification is applicable to data where the dependent (target) variable is categorical, for example the risk of a heart attack is either high, medium or low. First a training set, which comprises the already classified target variable along with its independent variables, is used to 'learn' from data, meaning which combination of variables is associated with which class. The model then sifts through the unclassified data set and classifies each entry according to a certain group (Larose, 2005).

Estimation builds a model by using a complete data set, using dependent and independent data points, similar to classification, except that the independent variable is numerical. New observations are added and the value of the dependent variable is then estimated from the provided independent (predictor) variables. For example, the systolic blood pressure can be estimated from a patient's gender, body mass index, age and blood-sodium levels (Larose, 2005). The estimation model will be based on the relationship between the predictor variables from the training set and the corresponding systolic blood pressure.

Prediction varies from estimation and classification with the results of the model being applicable to the future. For example, prediction is used to predict the stock price three months from now or the increase in the amount of deaths from traffic accidents if the speed limit is increased. Classification and estimation techniques are often applied for prediction (Larose, 2005).

Clustering is a method used to group similar data points into classes. A collection of observations that are similar is referred to as a cluster. With clustering, there is no independent variable, it does not classify, predict or estimate, but rather segments the data, maximising the similarity of data points within the cluster and minimising the similarity with regard to

other clusters. Clustering output is often used as input to further data mining using k-means or hierarchical clustering (Larose, 2005).

Association is used to find which attributes of data instances go together. The most common example is 'market-based analysis' which employs association to determine rules that describe the relationship between two or more attributes (Larose, 2005). For example, a supermarket may find that 80% of their shoppers who buy bread on a Sunday, also buy milk.

B.2 Regression analysis

This section presents additional information that was introduced in Section 3.6.

B.2.1 Fitting a model from data points

A straight line graph with two axes are constructed from two pieces of information, for example a slope and a point. A line is a mathematical certainty and not influenced by variation. Thus, when estimating a point and slope form data which has variation, it is termed as 'fitting a line'. Most straight line models use either the intercept and slope or the mean and slope (Riffenburgh, 2012). The slope-intercept model will have the form $y = \mathcal{B}_0 + \mathcal{B}_1 x$, where the intercept is $\mathcal{B}_0$ and the slope is represented by $\mathcal{B}_1$. The other method pertains to using the mean of both the variables, namely $\bar{x}$ and $\bar{y}$, which would result in

$$y - \bar{y} = \mathcal{B}_1(x - \bar{x}). \tag{B.1}$$

The best fit for the slope is usually calculated with the method of least squares, which entails finding the smallest sum of squares of the vertical distances between the data points and fitted line (Riffenburgh, 2012).

B.2.2 Confidence intervals for the regression model

The predictive ability of the model is indicated by the coefficient of determination R^2, which is generally used in models with one x predicting y by a straight line. R^2 evaluates how well the model of the x-and-y relationship predicts y as discussed in Section 3.6.1. In some cases, a confidence interval (CI) on the slope of the regression line is informative, with a tight CI suggesting a strong relationship (Riffenburgh, 2012).

The CI on the population slope ($\mathcal{B}_1$) at n-2 degrees of freedom (dof) is

$$P[b_1 - t_{1-\alpha/2} s_b < \mathcal{B}_1 < b_1 + t_{1-\alpha/2} s_b] = 1 - \alpha. \tag{B.2}$$

The CI on the prediction strength of the most likely (or average) y for a given x, written as $\bar{y}|x$, is more often investigated in research, whereas in clinical practice, the prediction for a specific

patient (or individual) is more of interest. In the first case, the population mean μ for a certain x $(\mu|x)$ is estimated by $\bar{y}|x$ with a standard error calculated as

$$s_{\bar{y}|x} = s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)(s_x)^2}}, \tag{B.3}$$

and the CI calculated as

$$P[\bar{y}|x - t_{1-\alpha/2}s_{(\bar{y}|x)} < \mu|x < \bar{y}|x + t_{1-\alpha/2}s_{(\bar{y}|x)}] = 1-\alpha. \tag{B.4}$$

With an individual patient prediction of y from x, the standard error is calculated differently, which changes the CI for the estimated expected population value of y for a given x, i.e. $E(y|x)$. The standard error is calculated as

$$s_{(y|x)} = s_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)(s_x)^2}}, \tag{B.5}$$

with a $1-\alpha$ confidence interval at $n-2$ (dof) of

$$P[y|x - t_{(1-\alpha/2)}s_{(y|x)} < E(y)|xE(y|x) < y|x + t_{(1-\alpha/2)}s_{(y|x)}] = 1-\alpha. \tag{B.6}$$

One of the questions regression analysis aims to answer is how much the regression slopes of two estimated samples vary, and is calculated by

$$s_{b(1-2)} = \sqrt{\left(\frac{(n_1-2)s_{e,1}^2 + (n_2-2)s_{e,2}^2}{n_1+n_2-4}\right)\left(\frac{1}{(n_1-1)s_1^2} + \frac{1}{(n_2-1)s_2^2}\right)}, \tag{B.7}$$

and a test statistic of

$$t_{ts} = \frac{b_{1,1} - b_{1,2}}{s_{b,1-2}}. \tag{B.8}$$

B.2.3 Correlation analysis

The correlation coefficient (r) is defined between -1 and 1, with a value of zero indicating no relationship between x and y. Exactly -1 or 1 indicates a perfect predictive ability as mentioned in Section 3.6.1. Other than with regression, x and y are not thought of as dependent or independent variables and the goal is not to predict the one from the other, but merely to know how closely x and y are associated in a straight line relationship.

There are a few underlying assumptions pertaining to correlation, namely:

1. Errors in the data set are independent of each other;
2. The data points have a linear relationship (r=0 indicates that a linear element is non-existent);
3. Exact readings on one axis is not required (both x and y may be measured with random variability); and
4. x and y follow a bivariate normal distribution, where x and y are the width and length, with the probability of any joint value of x and y making up for the height.

The correlation between two continuous variables is most commonly known as Pearson's correlation and in the case where the data is not continuous but ranked, it is known as Spearman's (or rank) correlation. In the case of ranked data, the assumption of bivariate normal distribution does not apply. The correlation coefficient is typically calculated using

$$r = \frac{s_{xy}}{s_x s_y}, \tag{B.9}$$

or, if regression was already calculated

$$r = b_1 \frac{s_x}{s_y}. \tag{B.10}$$

Rank correlation is calculated using

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{B.11}$$

where d_i is the x-rank minus the y-rank for patient i. Similar to regression, various tests including the CI for correlation as well as the population correlation coefficient ρ can be calculated. A significance test for $p = 0$ ($H_0 : \rho = 0$) can be determined at $n_s - 2$ *dof* as well as a test for ρ equal to some theoretical coefficient ($H_0 : \rho = \rho_0$). A z-test can be used with the case of comparing correlation coefficients of two samples (r_1 and r_2) (Riffenburgh, 2012).

B.2.4 PROC PHREG method for estimating 30-day readmission

This section presents more information on the study that modelled the time to readmission by using the SAS/STAT PHREG procedure as introduced in Section 3.6.4.3. The time-varying predictors included the number of previous readmissions, inpatient days and procedures performed at each admission. The fixed-value covariates were, for example, age, disabled status and the number of emergency room visits prior to initial hospitalisation (Lemke, 2012).

There are two possible methods used by PROC PHREG that incorporate time-dependent variables, namely a counting process and programming statements, which produce the same results. The counting process will be discussed briefly (Powell & Bagnell, 2012). A patient with two readmissions and a censoring time will have three time intervals in the analysis, each interval comprising a start time (t_{start}), an event time (t_{stop}) and an indicator of an readmission

(event=1) or censoring (event=0). At the first interval t_{start} is equal to zero, whereas at the next interval it is equal to the previous readmission time. A two-time-readmitted patient's first two intervals will have events equal to one, with the third event being censored or the end of the observation period being censored (Lemke, 2012).

When using the count process, the assumption for proportional hazard cannot be checked. The most popular methods for checking it are the scaled Schoenfeld residual and Cumulative score residual. If a model predicts equally well at all risk levels, the deviance residuals are not related to the risk scores. Furthermore, in the case of certain risk levels having a concentration of high deviance residuals, a poor fit is suggested (Lemke, 2012).

Recurrent readmission events and time-dependent covariates, which may change with each readmission, are included in the model. Cox regression incorporates time-varying covariates, given that each patient's covariate values are available every time anyone in the sample has an event (Lemke, 2012). The model treats each patient as a multiple-event counting process, using the hazard function

$$h(t|M) = h_0(t_k - t_{k-1})exp(m_i). \tag{B.12}$$

Here k is the current readmission and m a predictor variable. The counting process uses all the records for each patient, where a record corresponds to a time period during which the covariates are the same (Lemke, 2012). Time-varying variables can change value during the total time period. The logged hazard function, (3.10), is adapted to include time-varying variables (Powell & Bagnell, 2012). An example including one static and one time-dependent variable is

$$\log h_i(t) = \log h_0(t) + \beta_1 m_{i1} + \beta_2 m_{i2}(t). \tag{B.13}$$

Further reading on this subject:

- Powell & Bagnell (2012); and
- Lemke (2012).

B.2.5 Impurity: the Gini index and entropy

This section presents additional information about the Gini index and entropy function introduced in Section 3.7.1.1.

The concept of calculating the impurity function of a node to determine the best splits are mentioned in Section 3.7.1. Define τ as the parent node and τ_R and τ_L as the right and left daughter nodes respectively. Define $\Pi_1, ...\Pi_Q$ as the classes for Q $\geq$ two classes and $p(q|\tau)$ as a measure of the conditional probability that an observation $\mathbf{X}$ in Π_q falls into node τ, written as $P(\mathbf{X} \in \Pi_q|\tau)$.

Thus, for each node, the impurity function can be written as: $i(\tau) = \phi[p(1|\tau), ...p(Q|\tau)]$, where ϕ is a symmetric function defined on all probabilities $(p_1, ...p_q)$ with the sum maximised at points (1,0,...,0), (0,1,0,...,0) to (0,...,1). In the case where there are only two classes (Q=2), the

conditions reduce to a symmetric $\phi(p)$, which should be maximized at $p = \frac{1}{2}$ and $\phi(0) = \phi(1) = 0$. ϕ could thus be either the entropy function

$$i(\tau) = -\sum_{q=1}^{K} p(q|\tau) \log p(q|\tau) \tag{B.14a}$$

$$\text{and for Q=2:} \quad i(\tau) = -p \log p - (1-p) log(1-p), \tag{B.14b}$$

where p is set to $p = p(1|\tau)$ or the Gini diversity index

$$i(\tau) = \sum_{q \neq' q} p(q|\tau) p(q'|\tau) = 1 - \sum_{q} \{p(q|\tau)\}^2 \tag{B.15a}$$

$$\text{and for Q=2:} \quad i(\tau) = 2p(1-p). \tag{B.15b}$$

APPENDIX C

ANOVA analyses

This appendix contain mainly screenshots of the output pertaining to analysing the continuous variables with regard to readmission. The age, LOS and days discharge versus if a patient was readmitted or not are analysed by using ANOVA tests, the Mann-Whitney U non-parametric test and Cohen's effect size. The methods are discussed in Section 4.3.1. The results of the tests are discussed in the various sections of this document with the screenshots of the full output displayed in this appendix.

C.1 ANOVA: Age

The age of patients readmitted and not readmitted were analysed and is discussed in Section 5.1.1 with screenshots of the output of the tests displayed in this section. The output of the Mann-Whitney U test is displayed in Table C.1. The descriptive statistics for the age of patients readmitted an not readmitted after the various admissions are displayed in Table C.2. The results of Levene's test, which tests the assumption of equal variances, and graphs used to check normality are displayed respectively in Table C.3 and Figure C.1. The mean age for patients readmitted and not readmitted at each admission is displayed in Figure C.2.

	Mann-Whitney U Test (w/ continuity correction) (VirStatistica in Age box plots) By variable readmitted								
Variable	Rank Sum Group1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2
Age1	382447.0	893556.0	247740.0	1.689	**0.0912**	1.691	0.091	461	1136
Age2	38779.5	65873.5	23728.5	-0.611	**0.5411**	-0.612	0.541	173	284
Age3	7396.0	7482.0	3017.0	1.995	**0.0461**	1.997	0.046	78	94
Age4	1135.0	1946.0	574.0	-1.699	**0.0893**	-1.701	0.089	33	45
Age5	191.5	369.5	116.5	0.153	**0.8786**	0.153	0.878	11	22
Age6	17.0	49.0	11.0	-0.102	**0.9187**	-0.103	0.918	3	8

TABLE C.1: *Nonparametric test for difference between the mean age of patients readmitted or not readmitted after admission a.*

TABLE C.2: *Descriptive statistics and Cohen's effect size for the age of patients and readmission.*

Effect	Descriptive Statistics (VirStatistica in Age box plots) Level of Factor	N	Mean	Stdev	StdErr	-95%	+95%
Age1							
Total		1597	30.38760	9.395017	0.235096	29.92647	30.84873
readmitted1	0	1136	30.15493	9.327555	0.276744	29.61194	30.69792
readmitted2	1	461	30.96095	9.545235	0.444566	30.08732	31.83459
Effect size:	0,09(negligible)						
Age2							
Total		457	31.83589	9.541770	0.446345	30.95874	32.71303
readmitted2	1	173	31.38728	9.160409	0.696453	30.01259	32.76198
readmitted2	0	284	32.10915	9.772612	0.579898	30.96769	33.25062
Effect size:	0,08(negligible)						
Age3							
Total		172	31.68023	9.053904	0.690354	30.31752	33.04295
readmitted3	0	94	30.62766	9.029205	0.931291	28.77830	32.47702
readmitted3	1	78	32.94872	8.977457	1.016497	30.92461	34.97282
Effect size:	0,26(small)						
Age4							
Total		78	33.62821	9.069872	1.026961	31.58327	35.67315
readmitted4	1	33	31.39394	7.745722	1.348357	28.64743	34.14045
readmitted4	0	45	35.26667	9.686917	1.444040	32.35639	38.17694
Effect size:	0,44(medium)						
Age5							
Total		33	31.72727	7.722826	1.344371	28.98888	34.46567
readmitted5	1	11	31.90909	7.091608	2.138200	27.14488	36.67330
readmitted5	0	22	31.63636	8.179798	1.743939	28.00964	35.26308
Effect size:	0,04(negligible)						
Age6							
Total		11	31.90909	7.091608	2.138200	27.14488	36.67330
readmitted6	1	3	31.00000	5.000000	2.886751	18.57931	43.42069
readmitted6	0	8	32.25000	8.013382	2.833158	25.55065	38.94935
Effect size:	0,19(small)						

TABLE C.3: *Test the assumption of equal variance for ages at admission a.*

	Levene's Test for Homogeneity of Variances			
	MS Effect	**MS Error**	**F**	**p**
Age1	13.588	31.134	0.436	0.509
Age2	36.836	32.190	1.144	0.285
Age3	19.950	30.585	0.652	0.420
Age4	58.330	23.133	2.522	0.116
Age5	7.333	17.904	0.410	0.527
Age6	21.879	14.241	1.536	0.247

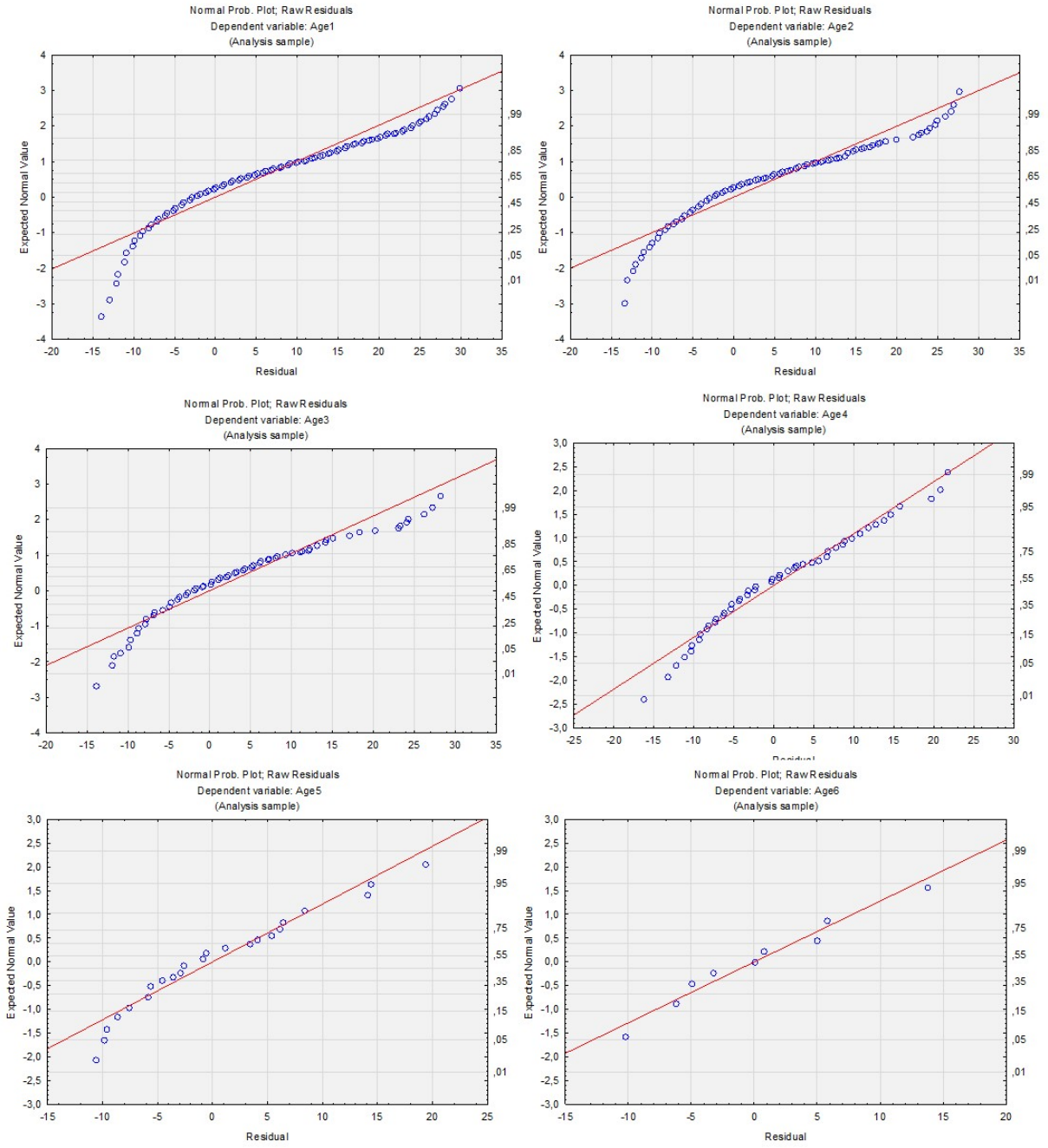


FIGURE C.1: *Testing assumption of normality for ages at admission a.*

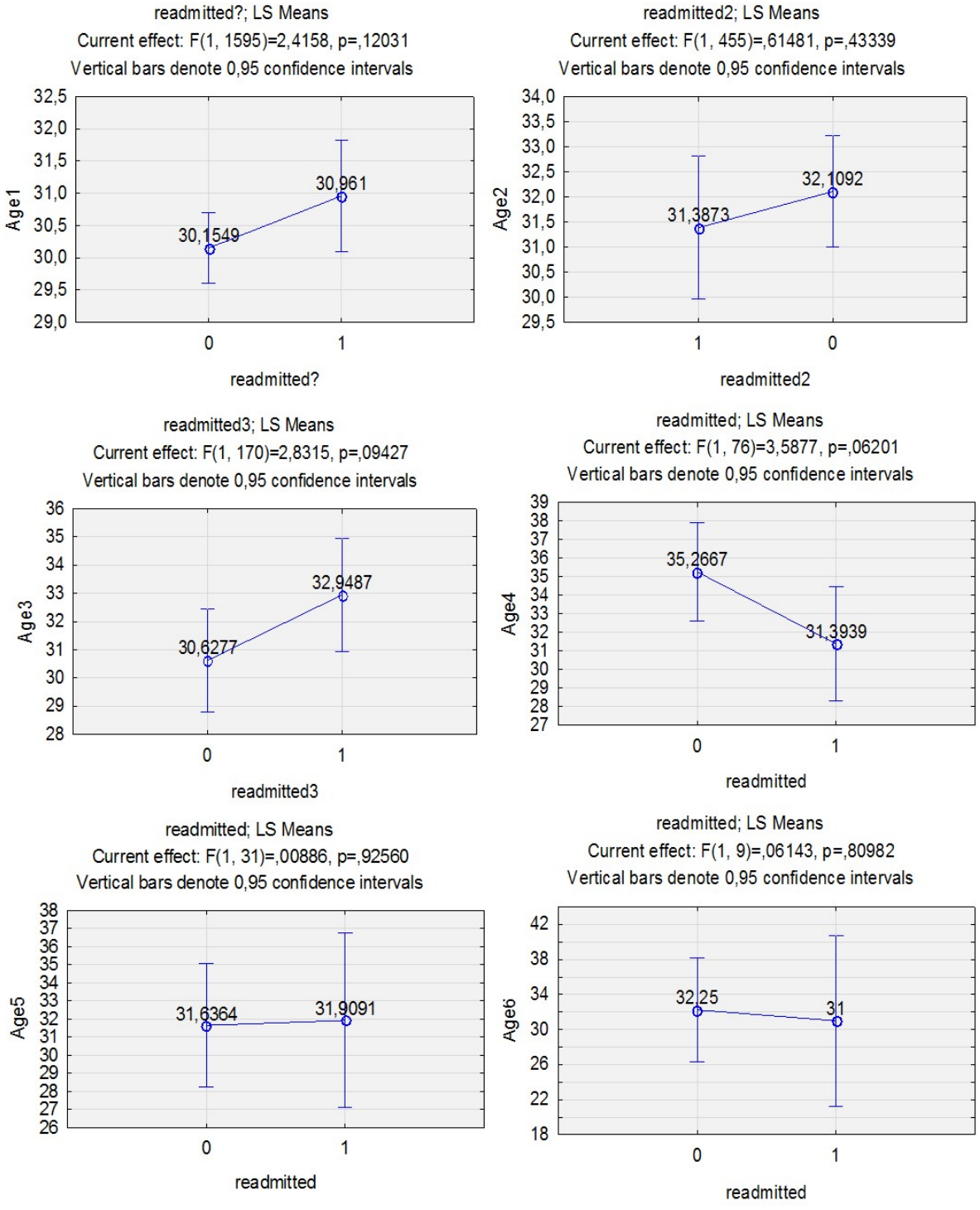


FIGURE C.2: *Box plots for patients' age at each admission.*

C.2 ANOVA: Length of stay

The LOS of patients readmitted and not readmitted were analysed and is discussed in Section 5.1.2. The descriptive statistics for the LOS after the various admissions are displayed in Figure C.4 with the results of Levene's test and the graphs used to check normality displayed respectively in Table C.5 and Figure C.4. The output of the Mann-Whitney U test is displayed in Table C.3. The average LOS for patients readmitted and not readmitted displayed in Figure C.5.

Effect	Descriptive Statistics (VirStatistica in Age box plots) Level of Factor	N	Mean	Stdev	StdErr	-95%	+95%
LOS1							
Total		1597	45.301	48.070	1.203	42.941	47.660
readmitted1	0	1136	44.560	47.951	1.423	41.768	47.351
readmitted1	1	461	47.126	48.365	2.253	42.699	51.552
Effect size:	0,05(negligible)						
LOS2							
Total		457	50.322	77.156	3.609	43.229	57.414
readmitted2	1	173	38.173	30.993	2.356	33.522	42.824
readmitted2	0	284	57.722	94.145	5.586	46.725	68.718
Effect size:	0,26(small)						
LOS3							
Total		172	52.163	56.199	4.285	43.704	60.621
readmitted3	0	78	41.141	41.604	4.711	31.761	50.521
readmitted3	1	94	61.309	64.713	6.675	48.054	74.563
Effect size:	0,37(small)						
LOS4							
Total		78	59.462	73.253	8.294	42.946	75.977
readmitted4	0	33	52.545	61.096	10.635	30.882	74.209
readmitted4	1	45	64.533	81.324	12.123	40.101	88.966
Effect size:	0,17(small)						
LOS5							
Total		33	48.424	57.759	10.054	27.944	68.905
readmitted5	0	11	28.364	18.619	5.614	15.855	40.872
readmitted5	1	22	58.455	67.840	14.464	28.376	88.533
Effect size:	0,55(medium)						
LOS6							
Total		11	55.273	52.017	15.684	20.327	90.219
readmitted6	0	3	64.333	66.425	38.351	-100.676	229.343
readmitted6	1	8	51.875	50.561	17.876	9.605	94.145
Effect size:	0,25(small)						

TABLE C.4: *Descriptive statistics and Cohen's effect size for the LOS of patients and readmission.*

	Levene's Test for Homogeneity of Variances			
	MS Effect	**MS Error**	**F**	**p**
LOS1	55.815	1693.762	0.033	0.856
LOS2	68290.283	4469.649	15.279	0.000
LOS3	17712.186	1908.071	9.283	0.003
LOS4	6120.889	2911.460	2.102	0.151
LOS5	7728.116	1671.432	4.624	0.039
LOS6	330.013	759.025	0.435	0.526

TABLE C.5: *Test the assumption of equal variance for LOS at admission a.*

Variable	Mann-Whitney U Test (w/ continuity correction) (VirStatistica in Age box plots) By variable readmitted: Rank Sum Group1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2
LOS1	392210.5	883792.5	237976.5	2.858	**0.0043**	2.859	0.004	461	1136
LOS2	37463.0	67190.0	22412.0	-1.573	**0.1158**	-1.573	0.116	173	284
LOS3	5936.0	8942.0	2855.0	-2.493	**0.0127**	-2.494	0.013	78	94
LOS4	1272.0	1809.0	711.0	-0.314	**0.7539**	-0.314	0.754	33	45
LOS5	161.0	400.0	95.0	-0.974	**0.3301**	-0.974	0.330	11	22
LOS6	21.0	45.0	9.0	0.510	**0.6098**	0.510	0.610	3	8

FIGURE C.3: *Nonparametric test for difference between the mean LOS of patients readmitted or not readmitted after admission a.*

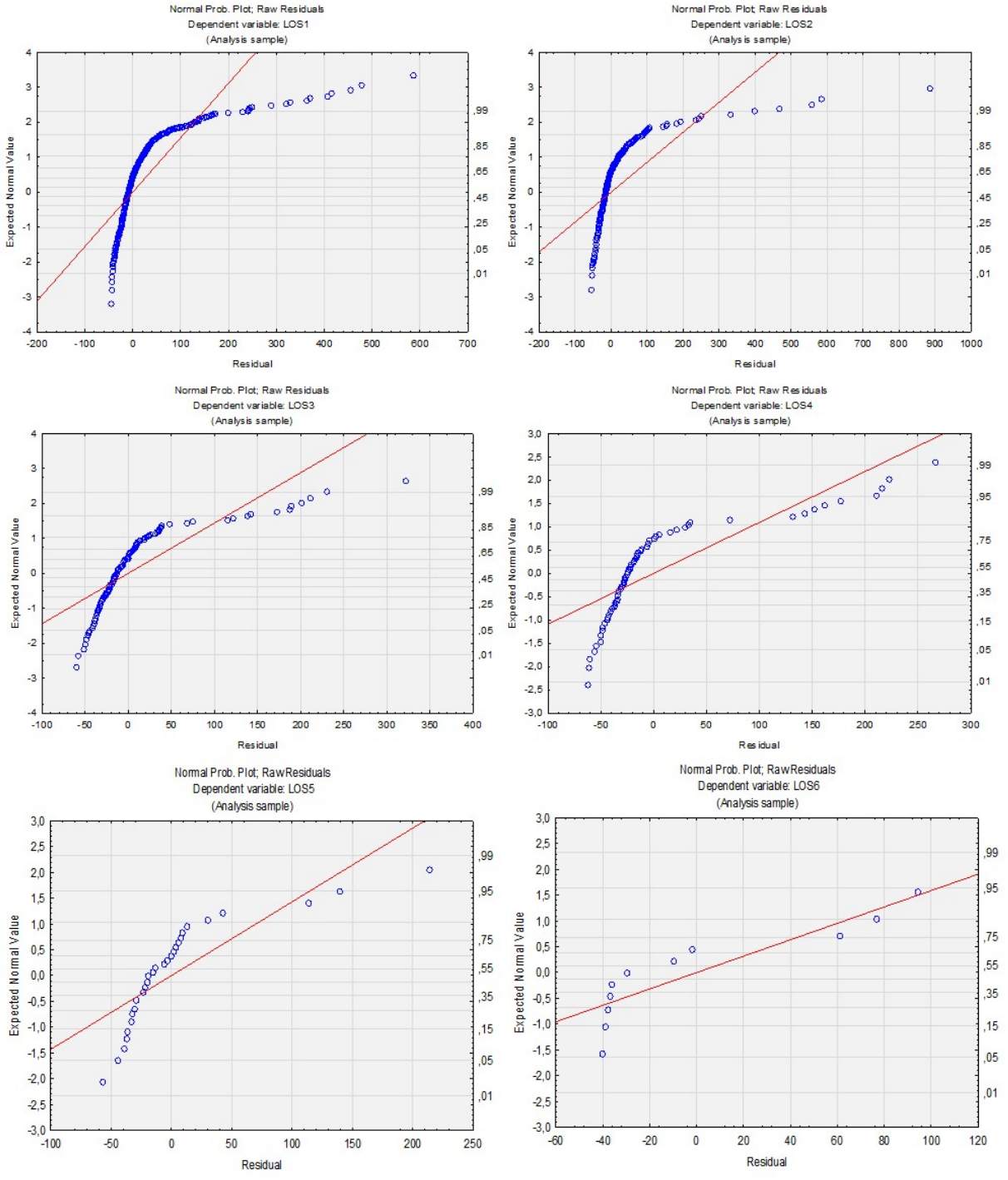


FIGURE C.4: *Testing assumption of normality for LOS at admission a.*

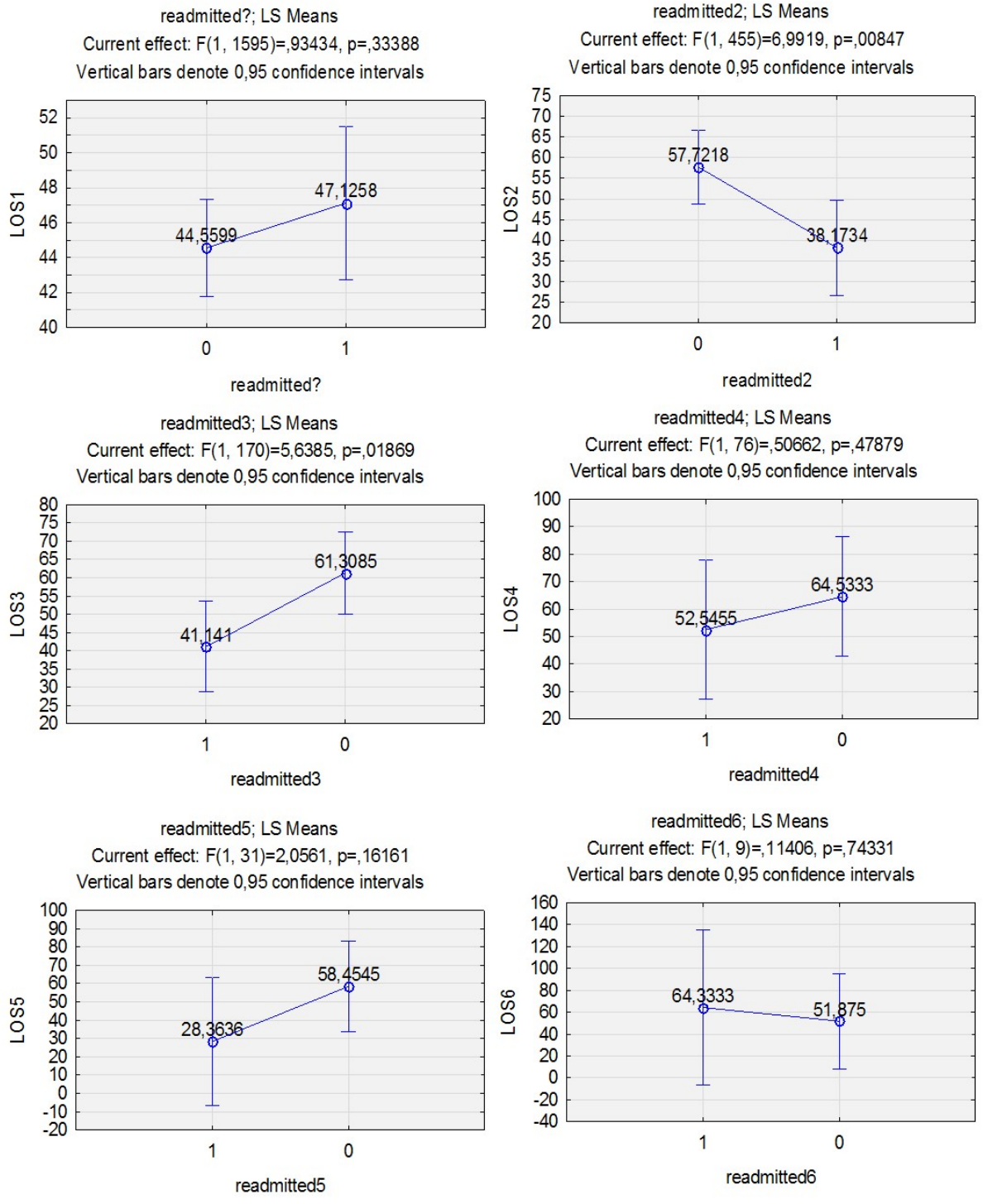


FIGURE C.5: *Box plots for patients' LOS at each admission.*

C.3 ANOVA: Days discharged

The amount of days patients were previously discharged before an admission and if they were readmitted are discussed in Section 5.1.3. The descriptive statistics are displayed in Table C.6 with the results of Levene's test and the graphs used to check normality displayed respectively in Table C.7 and Figure C.6. The output of the Mann-Whitney U test is displayed in Table C.8 and the graphs displaying the mean days discharged between patients readmitted and not readmitted are displayed in C.7.

TABLE C.6: *Descriptive statistics and Cohen's effect size for the days patients were discharged and readmission.*

Effect	Descriptive Statistics (VirStatistica in Age box plots) Level of Factor	N	Mean	Stdev	StdErr	-95%	+95%
DaysDischarged2							
Total		457	250.3260	220.3983	10.30980	230.0654	270.5866
readmitted2	1	173	215.4104	201.5601	15.32433	185.1624	245.6584
readmitted2	0	284	271.5951	228.8646	13.58062	244.8632	298.3269
Effect size:	0,25(small)						
DaysDischarged3							
Total		172	181.8953	177.5972	13.54166	155.1650	208.6257
readmitted3	0	78	159.8974	147.0556	16.65076	126.7415	193.0533
readmitted3	1	94	200.1489	198.3603	20.45930	159.5208	240.7770
Effect size:	0,23(small)						
DaysDischarged4							
Total		78	139.3462	129.1173	14.61965	110.2347	168.4576
readmitted4	0	33	115.7273	100.5607	17.50537	80.0700	151.3845
readmitted4	1	45	156.6667	145.2412	21.65128	113.0314	200.3019
Effect size:	0,32(small)						
DaysDischarged5							
Total		33	133.9394	133.2549	23.19670	86.6893	181.1895
readmitted5	0	11	117.2727	114.2262	34.44049	40.5345	194.0109
readmitted5	1	22	142.2727	143.6199	30.61987	78.5952	205.9502
Effect size:	0,19(small)						
DaysDischarged6							
Total		11	72.9091	69.8333	21.05553	25.9944	119.8237
readmitted6	0	3	40.0000	18.3303	10.58301	-5.5350	85.5350
readmitted6	1	8	85.2500	78.9462	27.91169	19.2493	151.2507
Effect size:	0,71(medium)						

TABLE C.7: *Test the assumption of equal variance for days discharge before admission a.*

	Levene's Test for Homogeneity of Variances			
	MS Effect	**MS Error**	**F**	**p**
DAYS2	64020.92	16379.365	3.909	0.049
DAYS3	64616.59	12317.154	5.246	0.023
DAYS4	17248.51	7179.787	2.402	0.125
DAYS5	717.66	8266.846	0.087	0.770
DAYS6	4866.75	1602.626	3.037	0.115

TABLE C.8: *Nonparametric test for the difference between the mean days patients readmitted and not readmitted are discharged before admission a.*

Variable	Mann-Whitney U Test (w/ continuity correction) (VirStatistica in Age box plots) By variable readmitted Rank Sum Group1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2
DAYS2	36276.0	68377.0	21225.0	-2.4394	**0.0147**	-2.4395	0.0147	173	284
DAYS3	6473.5	8404.5	3392.5	-0.8397	**0.4011**	-0.8397	0.4011	78	94
DAYS4	1172.0	1909.0	611.0	-1.3249	**0.1852**	-1.3249	0.1852	33	45
DAYS5	173.0	388.0	107.0	-0.5156	**0.6062**	-0.5156	0.6062	11	22
DAYS6	15.5	50.5	9.5	-0.4082	**0.6831**	-0.4092	0.6824	3	8

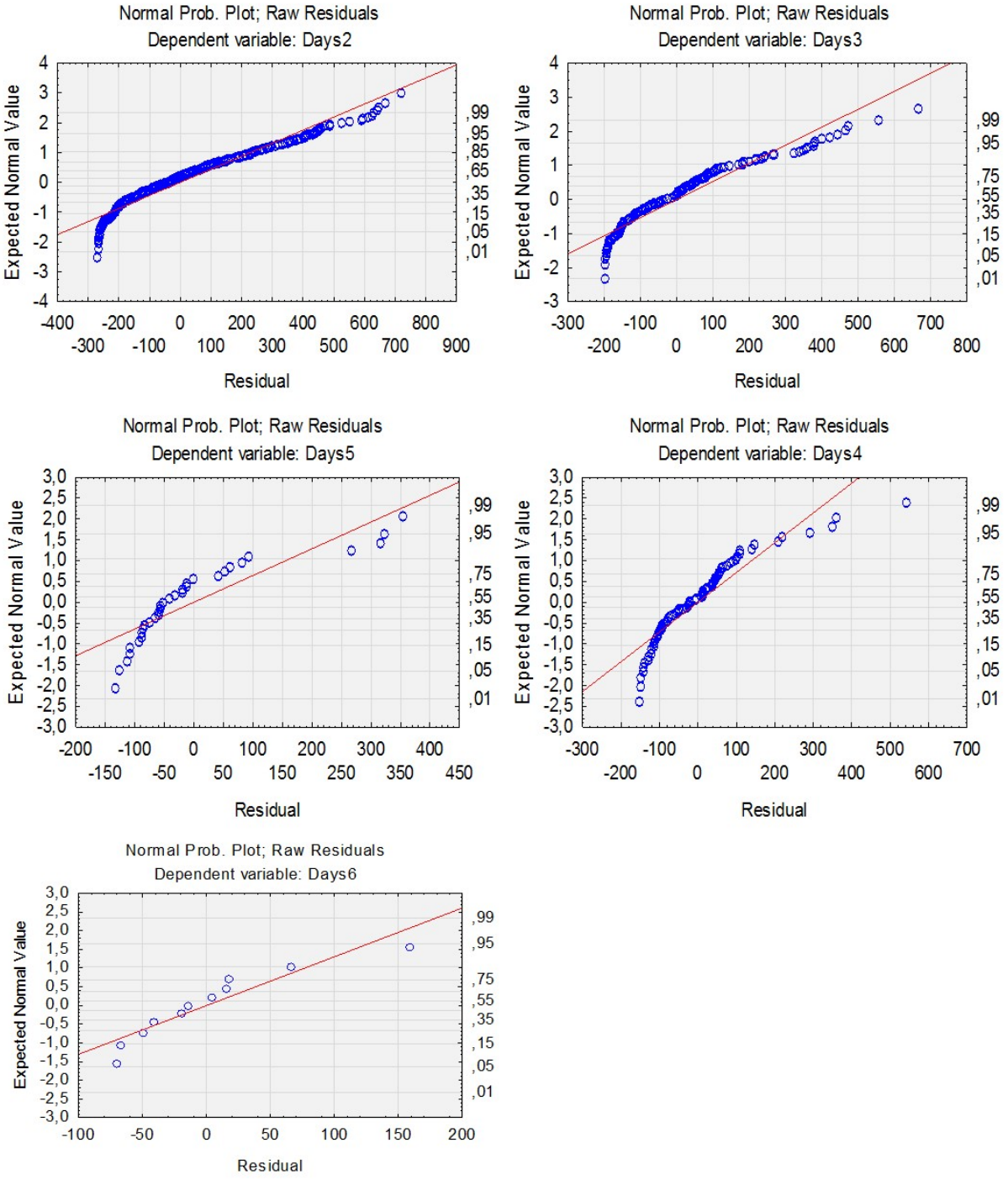


FIGURE C.6: *Testing assumption of normality for days discharge before admission a.*

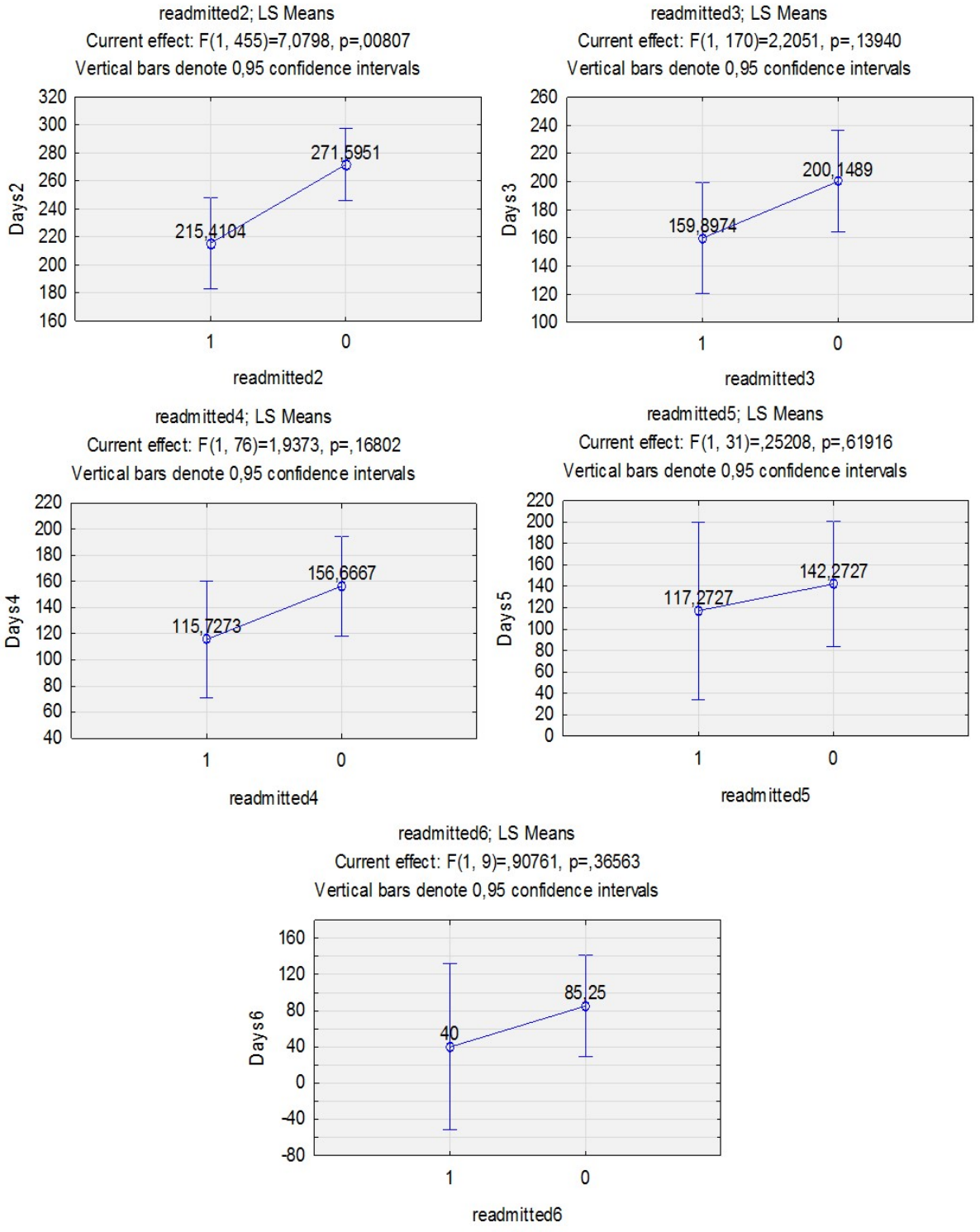


FIGURE C.7: *Mean days a patient was discharge before admission a and if he was readmitted thereafter.*

APPENDIX D

Additional descriptive analyses

This appendix pertains to analysing the variables separately with regard to readmission. Possible trends throughout the number of admissions in this study period are investigated in Section D.1. Chi-square tests conducted on the grouped dataset for the second and third admission data is presented in Section D.2.

D.1 Investigating possible trends throughout the various admissions

The variables are first individually analysed using descriptive methods which are discussed throughout Section 5.1. The percentage of observations per class in each variable throughout the admissions were investigated to determine if trends exist between the various admissions. The analyses are presented in this section.

Figure D.1 displays the percentage of patients admitted from a specific area for each of the admissions, but there is no apparent trend with the proportion of patients from an area staying rather constant. During the first three admissions the number of direct admissions (Area 4) increased slightly with patients for Paarl and surrounds (Area 1) decreasing slightly. The variation towards the last three admissions cannot be interpreted for the population owing to it only consisting of a small number of observations.

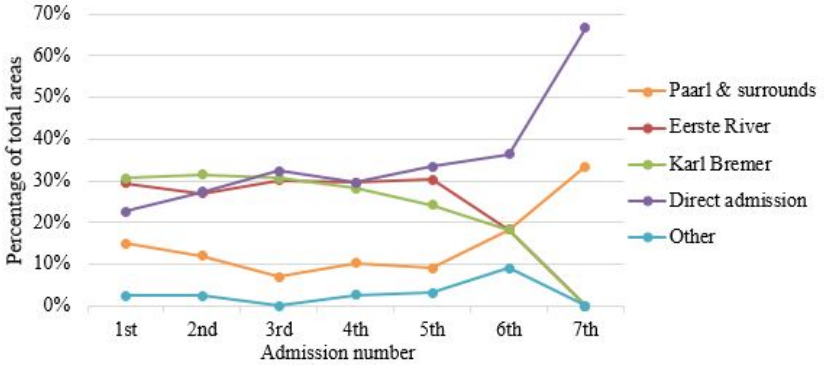


FIGURE D.1: *Percentage of patients admitted from a area throughout the admissions.*

The trends throughout the admissions for the diagnosis variable are displayed in D.2. The majority of the male patients admitted at Stikland have a primary diagnosis of schizophrenia. No apprent trends emerge between the admissions. SIPD diagnosis decreases from 20% to about 9% between the first two admissions, but stays constant after admission two. Throughout the admissions the prevalence of schizo-affective increases whilst bipolar stays more or less constant.

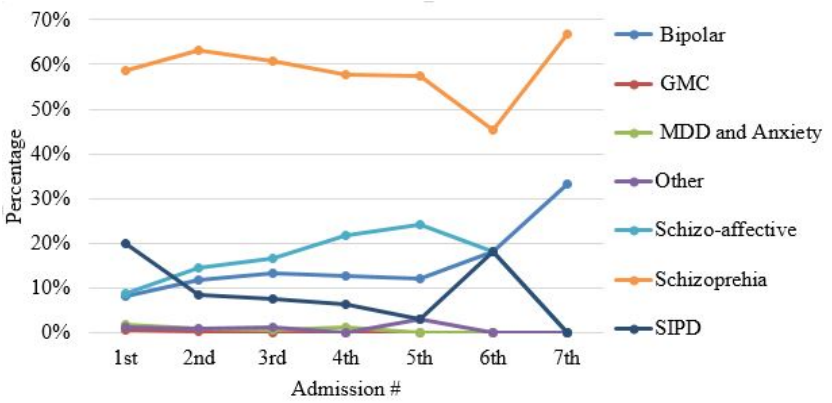


FIGURE D.2: *Percentage patients diagnosed with a certain ICD10-diagnosis throughout the admissions.*

The distribution of the follow-up variable throughout the various admissions are displayed in Figure D.3. The *other* and *none* classes both stay rather constant, *none* being 7% of all the classes and both experiencing an increase at the fifth admission. The graph does not convey much information from an engineering point of view, however the clinical SMEs considered the case where the place of follow-up at first admission may not have an effect on readmission and only become more important at the readmissions (Koen & Smit, 2016b).

The follow-up at PHC decreases slightly after which there is a 26% drop between the fourth and fifth admission. Patients following up at ACT and New Beginnings also increases between the fourth and fifth admissions. Until the fourth admission (3rd readmission) all classes are rather constant, after which a trend starts to emerge. Conclusions cannot be drawn from these trends however, owing to the sample sizes decreasing rapidly for each readmission (with 78 observations at fourth admission and 33 at fifth, 11 at sixth and three at seventh).

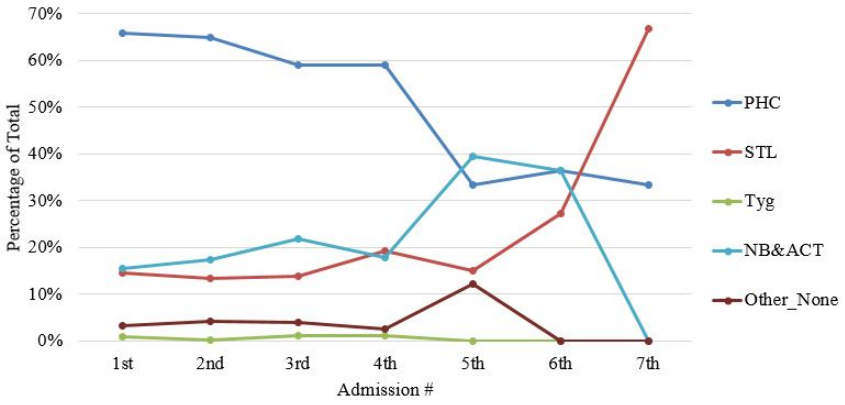


FIGURE D.3: *Percentage of patients following up at a certain place throughout the admissions.*

The percentage of patients belong to either ACT, New Beginnings or both throughout the admissions are displayed in Figure D.4. The percentage of patients at New Beginnings and ACT stay constant at between 1% and 3% for the first few admissions. The percentage of patients following up at New Beginnings also stay constant, around 14%, while the percentage

of ACT patients are less, being around 3% increasing to 18% at the fourth readmission. This sudden increase is most likely due to fewer observations. Not much inference can be made from the graph owing to fluctuations possibly caused by patients not being readmitted, patients starting to follow the programmes, or patients changing between the two or being subject to both.

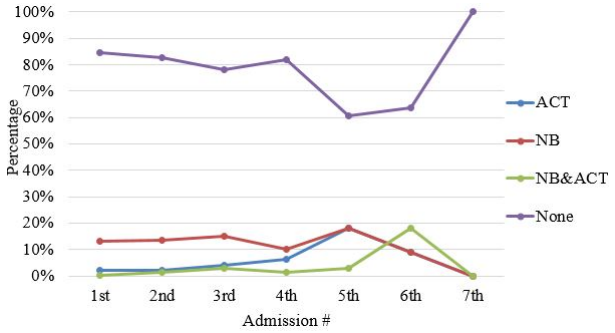


FIGURE D.4: *Percentage of patients joining a community care programme after admission a.*

D.2 Second and third admission data

The second admission and third admission data were analysed to determine whether a significant relationship exist between the various variables and readmission. It is realised that the results are applicable to this study period and may not be the patient's actual second or third admission, but until analysed, it is not clear if the admission data hold any significant information. The histograms are similar in shape to the first admission data.

D.2.1 Area

Figures D.5 to D.8 display the histograms and chi-square tests for area versus readmission for the second as well as the third admission data of this research. The results are discussed in Section 5.1.4.

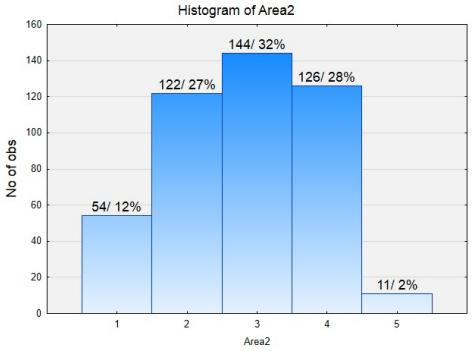


FIGURE D.5: *Histograms of the area-variable at the second admission.*

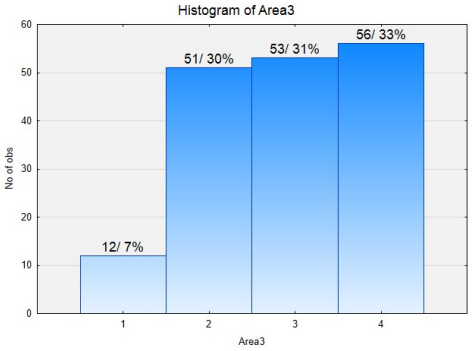


FIGURE D.6: *Histogram of the area-variable at the third admission.*

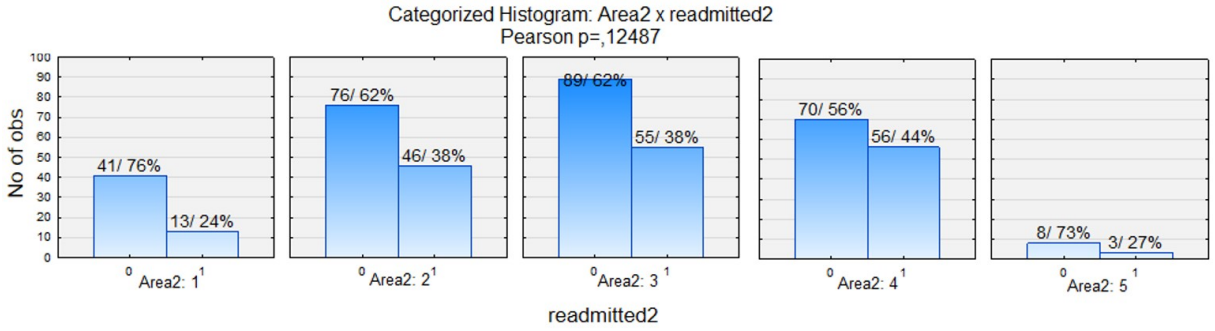


FIGURE D.7: *Categorised histogram of the area-variable at the second admission.*

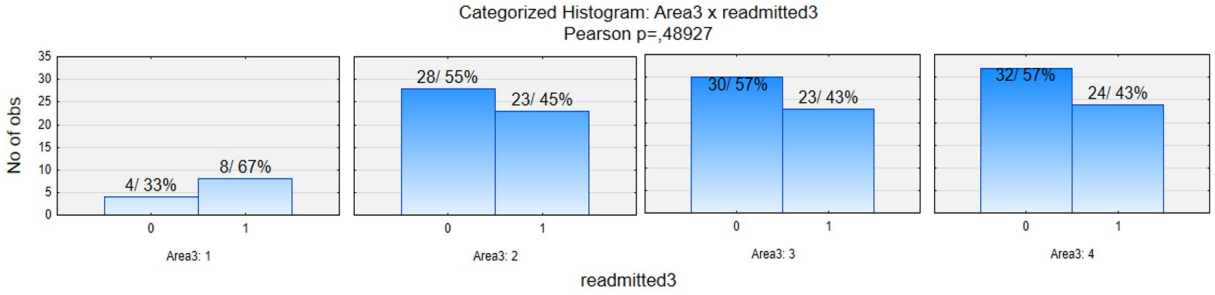


FIGURE D.8: *Categorised histogram of the area-variable at the third admission.*

D.2.2 Follow-up

Figures D.9 to D.12 display the histograms and chi-square tests for the follow-up variable versus readmission for the second as well as the third admission data of this research. These results are discussed in Section 5.1.6.

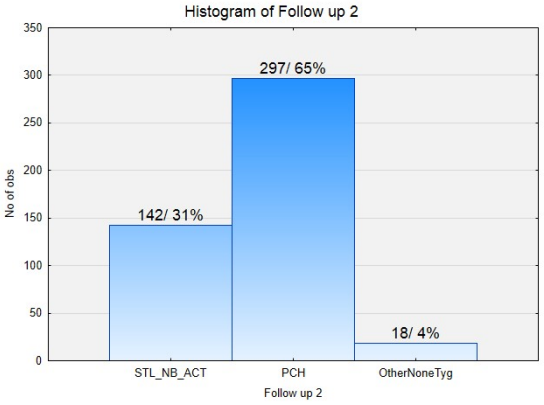


FIGURE D.9: *Histogram of the follow-up variable at the second admission.*

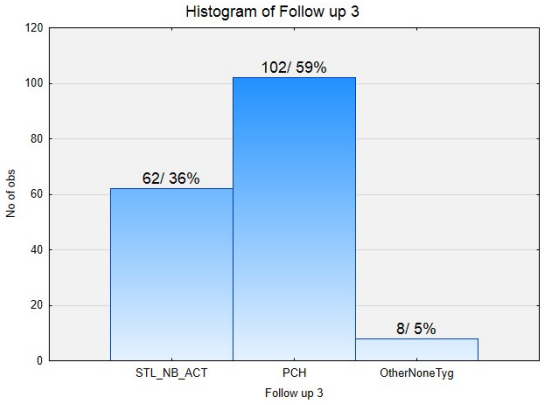


FIGURE D.10: *Histogram of the follow-up variable at the third admission.*

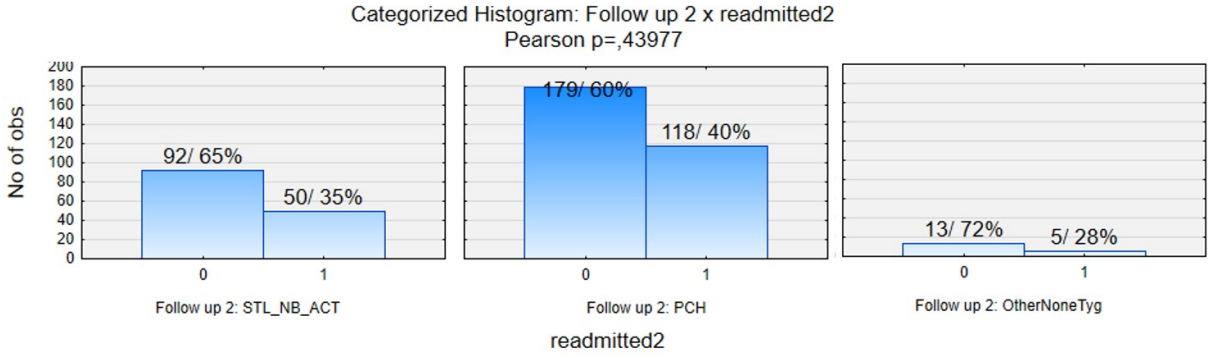


FIGURE D.11: *Categorised histogram of the follow-up variable at the second admission.*

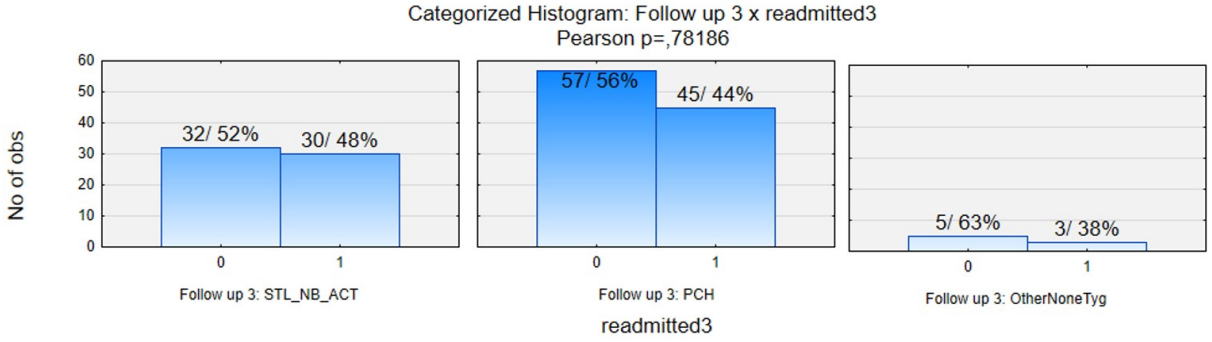


FIGURE D.12: *Categorised histogram of the follow-up variable at the third admission.*

D.2.3 ACT/NB

Figures D.13 to D.16 display the histograms and chi-square tests for the ACT/NB-variable versus readmission for the second as well as the third admission data of this research. These results are discussed in Section 5.1.7.

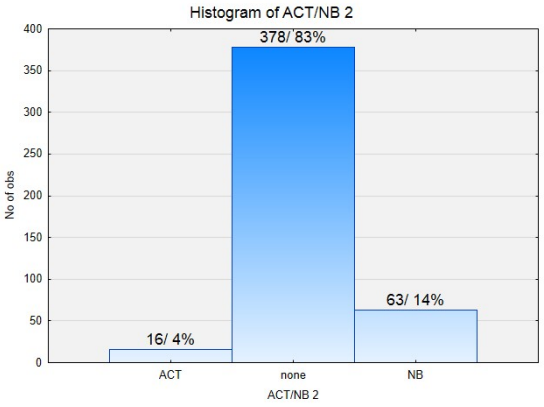


FIGURE D.13: *Histograms of the ACT/NB variable at the second admission.*

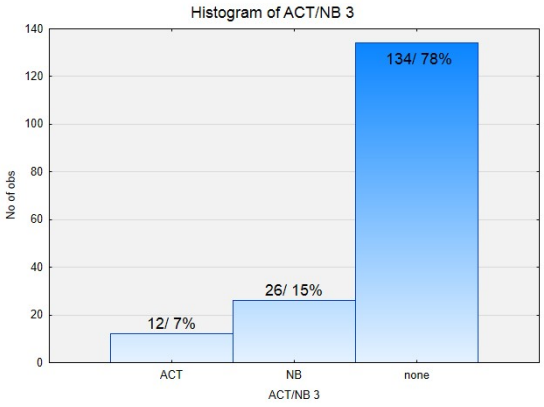


FIGURE D.14: *Histograms of the ACT/NB variable classes at the third admission.*

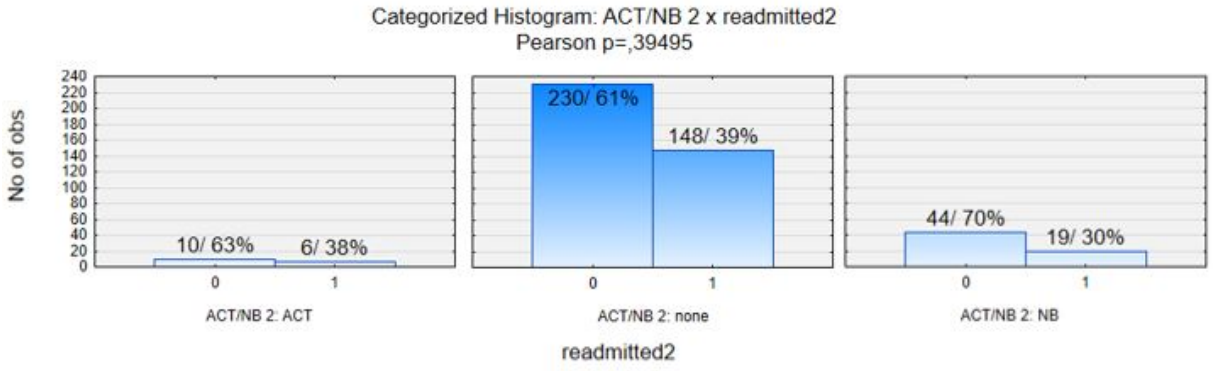


FIGURE D.15: *Categorised histogram of the ACT/NB variable at the second admission.*

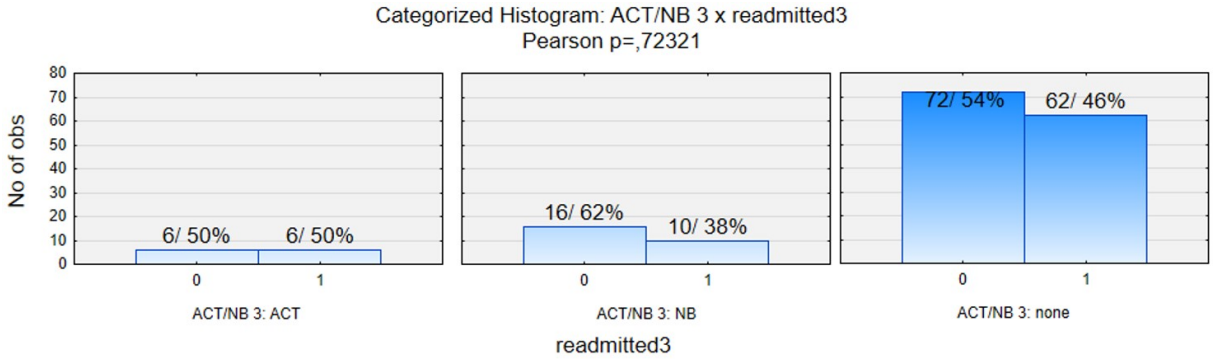


FIGURE D.16: *Categorised histogram of the ACT/NB variable at the third admission.*

D.2.4 ICD10 diagnosis

Figures D.17 to D.20 display the histograms and chi-square tests for the ICD10 diagnosis variable versus readmission for the second as well as the third admission data of this research. The results are discussed in Section 5.1.5.

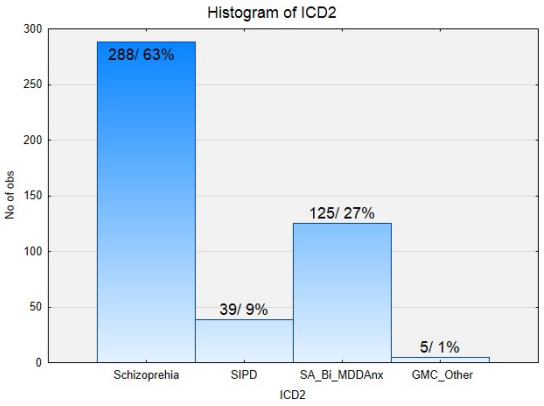


FIGURE D.17: *Histograms of the ICD10-diagnosis variable at the second admission.*

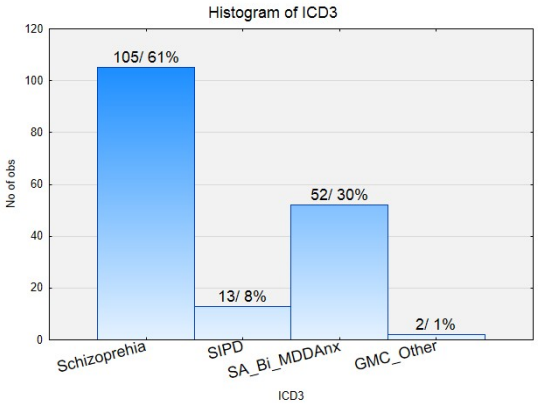


FIGURE D.18: *Histograms of the ICD10-diagnosis variable at the third admission.*

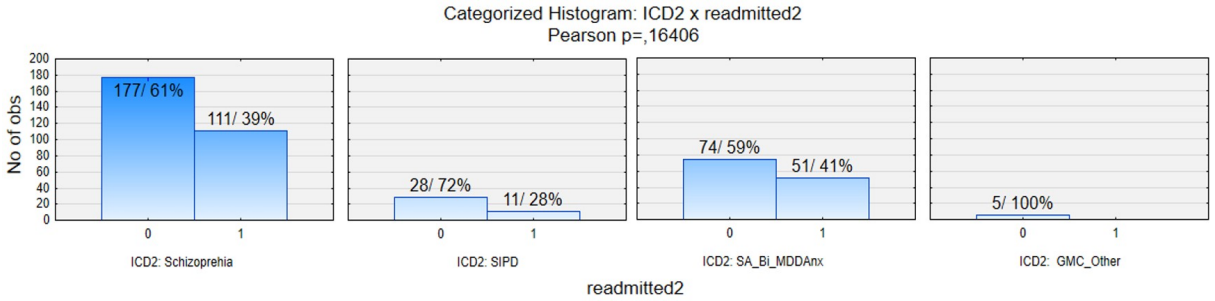


FIGURE D.19: *Categorised histogram of the ICD10-diagnosis variable at the second admission.*

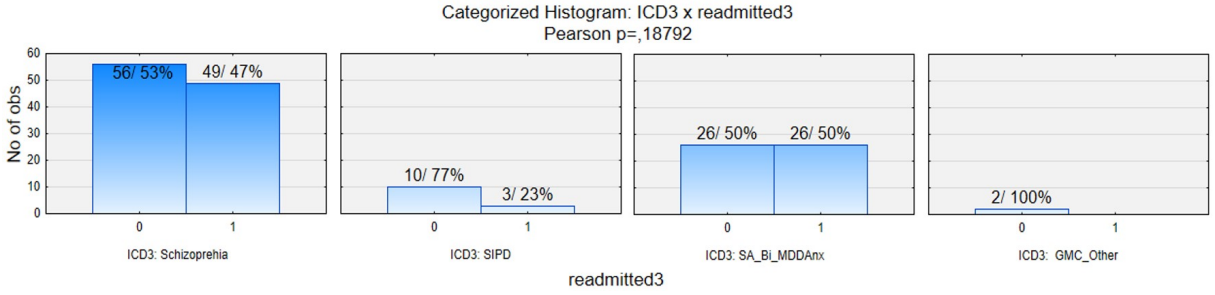


FIGURE D.20: *Categorised histogram of the ICD10-diagnosis variable at the third admission.*

APPENDIX E

Substance dataset

This appendix contains information about the various analyses that were conducted on the substance dataset. The dataset constitutes 297 entries of which the substance use is known to be accurate. The main goal of analysing this dataset was to determine if substance use is related to readmission and is discussed mainly in Section 5.1.8.

E.1 ANOVA analysis

The age, LOS and days discharged were analysed with ANOVA, the Mann-Whitney U test and Cohen's effect size with the results presented in this section.

E.1.1 Age

ANOVA analysis was conducted on the data with regard to readmission and compared to the primary dataset's findings. Figure E.1 displays the graphs that were used to check the normality assumption and Figure E.2 display the mean age of patients readmitted and not readmitted. Table E.1 and Table E.2 respectively display the results of the Mann-Whitney U test and Levene's test.

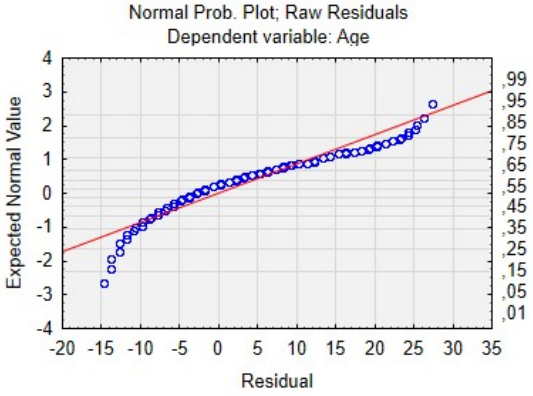


FIGURE E.1: *Check normality assumption for age (not satisfied).*

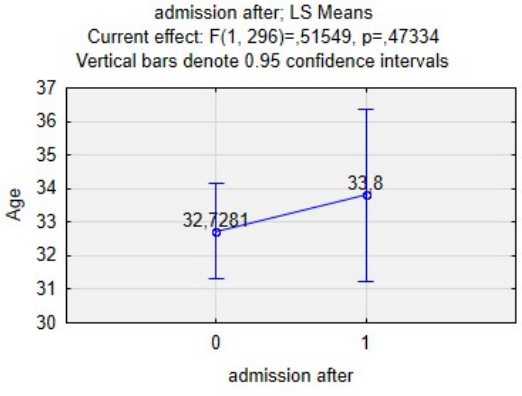


FIGURE E.2: *Difference in the mean age between patients readmitted and not readmitted.*

TABLE E.1: *Mann-Whitney U test comparing the mean age of patients readmitted or not readmitted.*

	Mann-Whitney U Test (w/ continuity correction)								
variable	Rank Sum	Rank Sum	U	Z	p-value	Z adjusted	p-value adjusted	Valid N	Valid N
Age	10756.0	33795.0	7689.0	0.4607	**0.6450**	0.4611	0.6448	70	228

TABLE E.2: *Test the assumption of equal variance for ages at admission a (satisfied).*

	Levene's Test for Homogeneity of Variances			
	MS	MS	F	p
Age	52.6785	39.5884	1.3307	**0.2496**

E.1.2 Length of stay

The LOS was analysed by ANOVA analysis and compared to the primary dataset's findings. Table E.3 and Table E.4 respectively display the results of the Mann-Whitney U test and Levene's test. Figure E.3 displays the test for normality and Figure E.4 displays the mean LOS of patients readmitted and not readmitted.

TABLE E.3: *Mann-Whitney U test comparing the mean length of stay of patients readmitted or not readmitted.*

	Mann-Whitney U Test (w/ continuity correction)								
variable	Rank Sum	Rank Sum	U	Z	p-value	Z adjusted	p-value adjusted	Valid N	Valid N
LOS	11106.5	33444.5	7338.5	1.0165	**0.3094**	1.0166	0.3093	70	228

TABLE E.4: *Test the assumption of equal variance for the LOS at admission a (satisfied).*

	Levene's Test for Homogeneity of Variances			
	MS	MS	F	p
LOS	0.7980	6176.2012	0.0001	**0.9909**

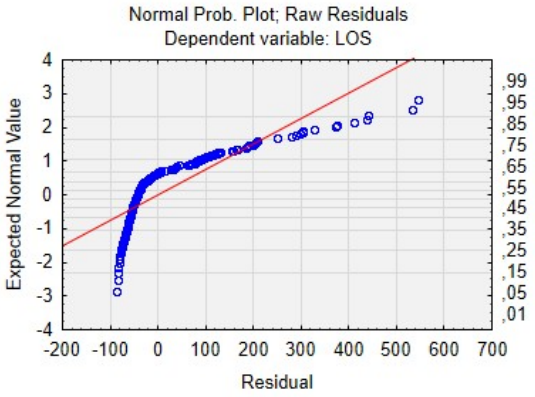


FIGURE E.3: *Check normality assumption for LOS (not satisfied).*

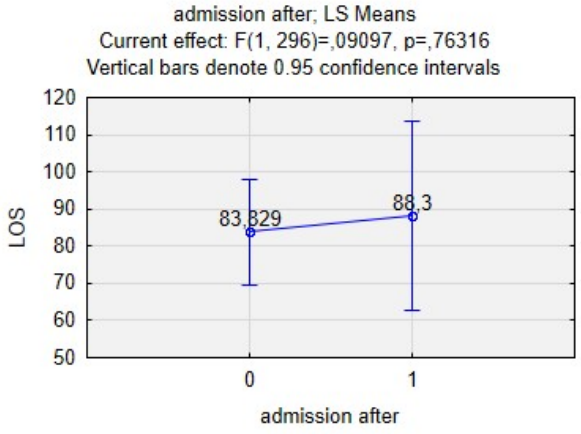


FIGURE E.4: *Difference in means of LOS between patients readmitted and not readmitted.*

E.1.3 Days discharged

Similarly to the age and LOS of patients the days a patient was discharge before a admission and then if they were readmitted were investigated. Table E.5 and Table E.6 respectively display the results of the Mann-Whitney U test and Levene's test. Figure E.5 displays the graphs that were used to check the normality assumption and Figure E.6 display the mean age of patients readmitted and not readmitted.

TABLE E.5: *Mann-Whitney U test comparing the mean days discharged between patients readmitted or not.*

variable	Mann-Whitney U Test (w/ continuity correction)								
	Rank Sum	Rank Sum	U	Z	p-value	Z adjusted	p-value adjusted	Valid N	Valid N
DaysDischarged	4453.5	2806.5	1531.5	1.1604	**0.2459**	1.1605	0.2459	70	50

TABLE E.6: *Test the assumption of equal variance for days discharged at admission a.*

	Levene's Test for Homogeneity of Variances			
	MS	MS	F	p
Days Discharged	5301.0807	23928.4414	0.2215	**0.6387**

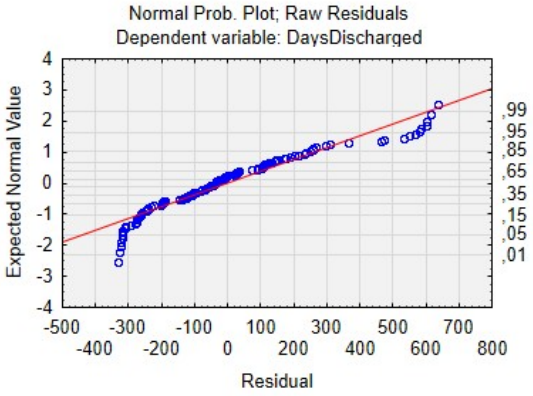


FIGURE E.5: *Check normality assumption for days discharged.*

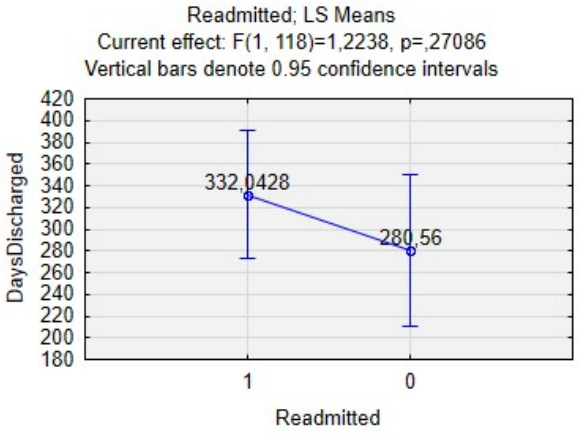


FIGURE E.6: *Difference in means of days discharged between patients readmitted and not readmitted.*

E.2 Histograms of the variables in the substance dataset

The histograms of the variables in the substance dataset are displayed in Figures E.7 to E.10. This was briefly compared to that of the larger dataset predominantly used for this research.

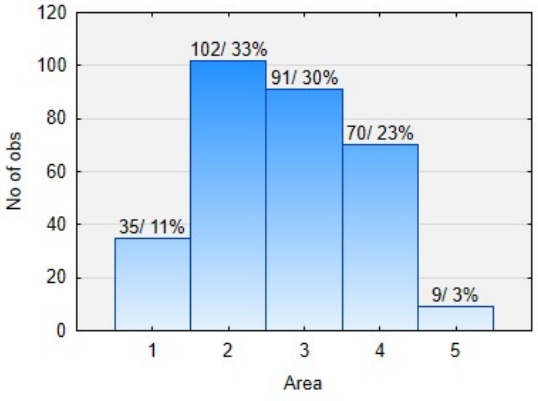


FIGURE E.7: *Histogram of the area variable*

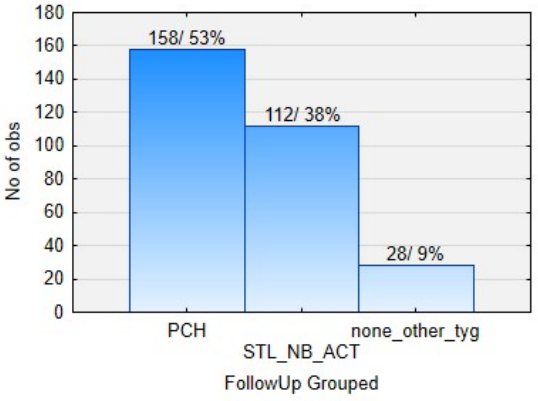


FIGURE E.8: *Histogram of the follow-up variable*

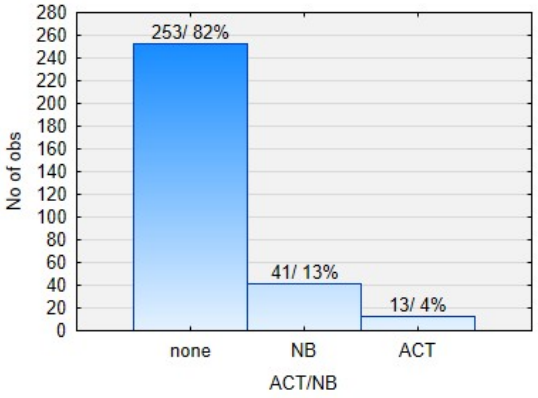


FIGURE E.9: *Histogram of the ACT/NB variable*

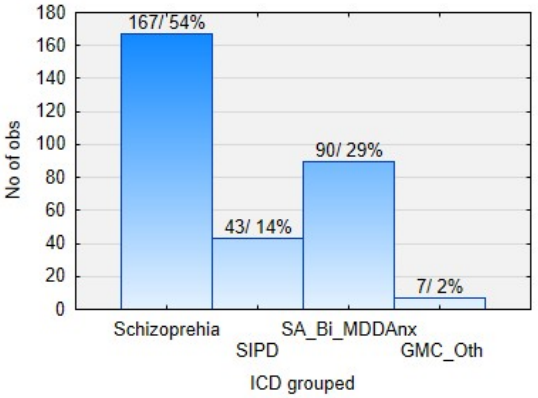


FIGURE E.10: *Histogram of the diagnosis variable*

E.3 Chi-square tests

Similarly to analysing the large dataset the categorical variables of the substance dataset were analysed to determine mainly if substance abuse can be linked to readmission. The results are compared to those of the larger dataset. The chi-square tests of the substance variables are displayed in Section E.3.1 and the other categorical variables in E.3.2.

E.3.1 Substance use

Figures E.11 to E.15 display categorised histograms from the chi-square analysis of the various substance abuse variables versus readmission.

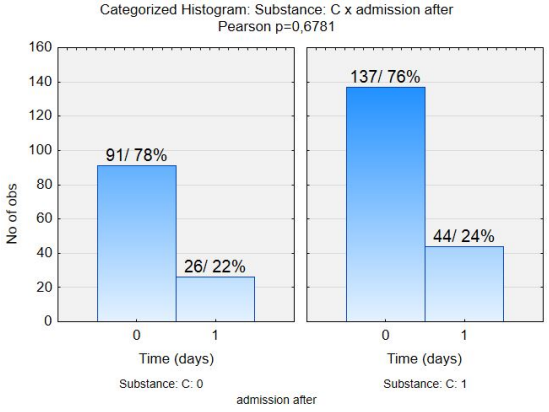


FIGURE E.11: *Chi-square test results for cannabis abuse and readmission.*

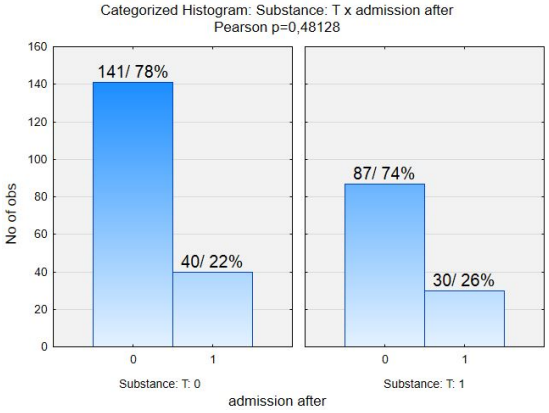


FIGURE E.12: *Chi-square test results for tik abuse and readmission.*

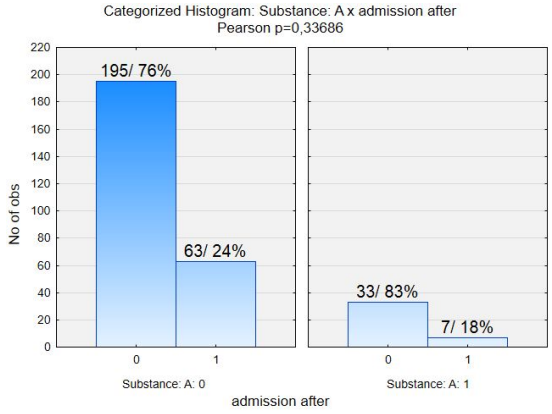


FIGURE E.13: *Chi-square test results for alcohol abuse and readmission.*

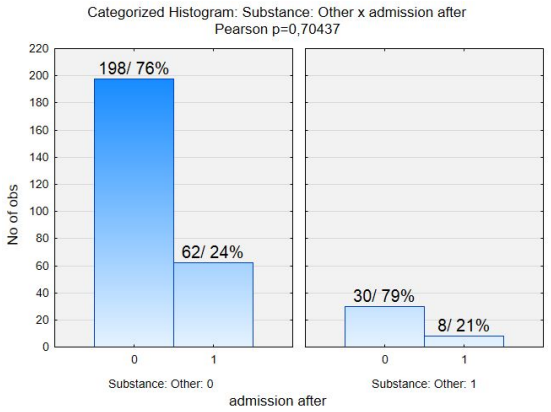


FIGURE E.14: *Chi-square test results for other-type substance abuse and readmission.*

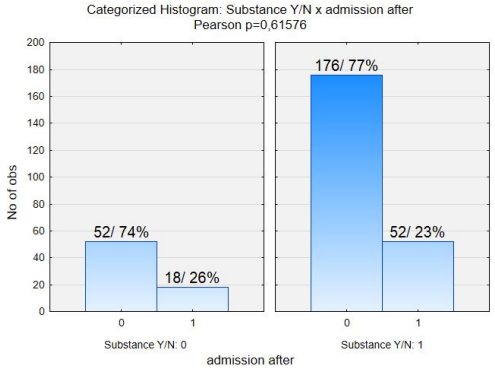


FIGURE E.15: *Chi-square test results for no substance abuse and readmission*

E.3.2 Area, follow-up, diagnosis and ACT/NB

The area, follow-up-, diagnosis- and ACT/NB variables were also briefly analysed to investigate if trends similar to that of the larger dataset exist and if there may be any other significant relationships between a variable and readmission. The results are displayed in Figures E.16 to E.19.

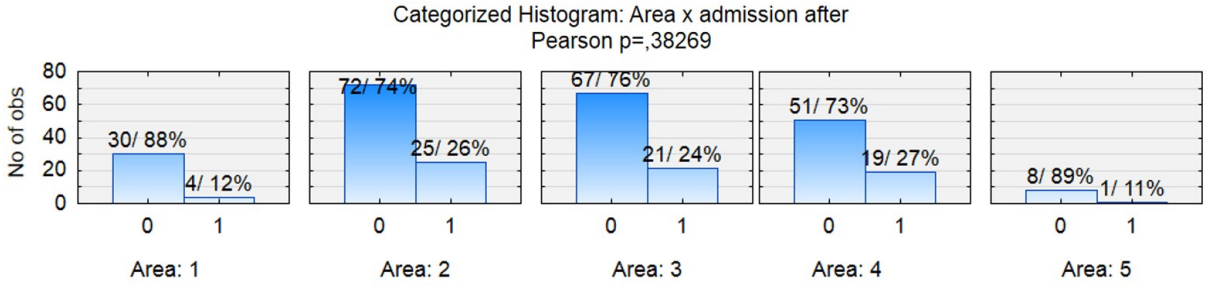


FIGURE E.16: *Chi-square test results for the area variable and readmission*

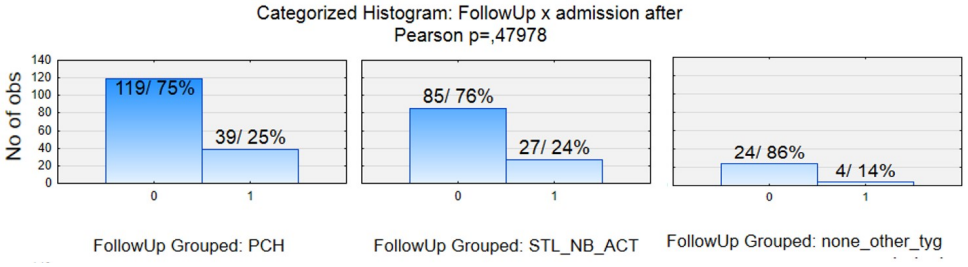


FIGURE E.17: *Chi-square test results for the follow-up variable and readmission*

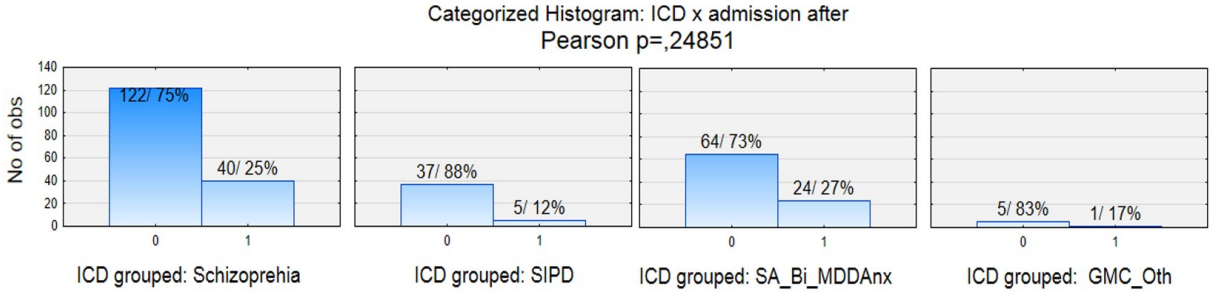


FIGURE E.18: *Chi-square test results for the ICD10-diagnosis variable and readmission*

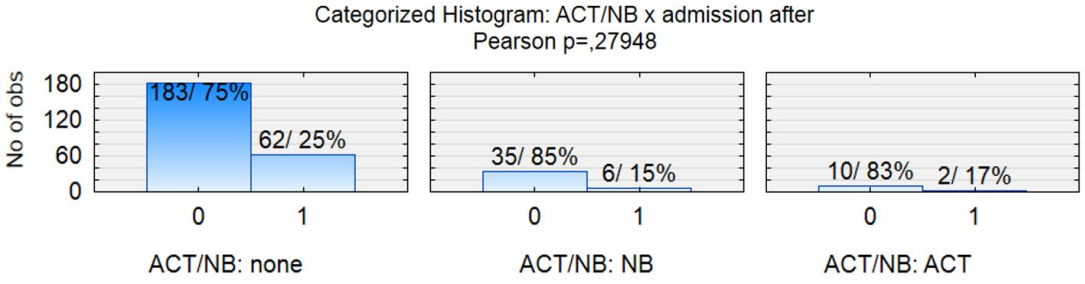


FIGURE E.19: *Chi-square test results for the ACT/NB variable and readmission*

APPENDIX F

Logistic regression and discriminant analysis

Logistic regression and discriminant analysis are used to determine variables that play a significant role in predicting readmission, with logistic regression generating more detailed results such as odds ratios. Both models are introduced in Section 4.3.1 and the results presented in Section 5.2.1.

F.1 Grouped dataset

Logistic regression and discriminant analysis were conducted on the grouped dataset with the results described in Section 5.2.1.1. This section presents the output of both the logistic regression model and discriminant analysis.

F.1.1 Logistic regression

As discussed in Section 4.3.1 the output for the logistic regression model of the grouped dataset include: the test of all effects (Table F.2), parameter estimates (Table F.1), odds ratios (Table F.4), ROC curve (Figure F.1) and the classification matrix (Table F.3) all displayed in this section.

TABLE F.1: *Parameter estimates for the grouped dataset.*

Effect	readmitted? - Parameter estimates (VirStatistica in Admi1Grouped) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1 Level of Effect	Column	**Estimate**	Standard Error	Wald Stat	Lower CL 95%	Upper CL 95%	p
Intercept		1	**-1.0016**	0.296	11.411	-1.583	-0.420	**0.0007**
Age1		2	**-0.0004**	0.006	0.003	-0.013	0.012	**0.9537**
LOS1		3	**0.0000**	0.001	0.001	-0.002	0.002	**0.9736**
Area1	5	4	**-0.1735**	0.302	0.329	-0.766	0.419	**0.5660**
Area1	1	5	**-0.1967**	0.148	1.767	-0.487	0.093	**0.1838**
Area1	2	6	**-0.0553**	0.121	0.209	-0.292	0.182	**0.6473**
Area1	3	7	**0.0854**	0.119	0.511	-0.149	0.319	**0.4746**
Follow up 1	PCH	8	**0.1060**	0.119	0.800	-0.126	0.338	**0.3711**
Follow up 1	STL_NB_ACT	9	**-0.1262**	0.142	0.795	-0.404	0.151	**0.3727**
Instit1	none	10	**-0.2213**	0.164	1.830	-0.542	0.099	**0.1762**
Instit1	NB	11	**-0.3390**	0.167	4.118	-0.666	-0.012	**0.0424**
ICD1	Schizophrenia	12	**0.3101**	0.143	4.689	0.029	0.591	**0.0303**
ICD1	SA_Bi_MDDAnx	13	**0.6649**	0.159	17.592	0.354	0.976	**0.0000**
ICD1	GMC_Other	14	**-0.4783**	0.373	1.645	-1.209	0.253	**0.1996**
Scale			**1.0000**	0.000		1.000	1.000	
Run 2:								
Area1	4	7	0.2976	0.1013	8.6398	0.0992	0.4961	0.0033
Follow up 1	OtherNoneTyg	8	**0.0202**	0.199	0.010	-0.370	0.410	**0.9191**
Instit1	ACT	11	**0.5603**	0.239	5.489	0.092	1.029	**0.0191**
ICD1	SIPD	12	**-0.4967**	0.173	8.256	-0.836	-0.158	**0.0041**

TABLE F.2: *Test of all effects for the grouped dataset.*

Effect	readmitted? - Test of all effects [LOGIT] Modeled probability readmitted? = 1 Degr. of freedom	Wald Stat	p
Intercept	1	11.411	0.000730
Age1	1	0.003	0.953659
LOS1	1	0.001	0.973597
Area1	4	10.371	0.034627
Follow up 1	2	1.908	0.385239
Instit1	2	5.707	0.057656
ICD1	3	37.447	0.000000

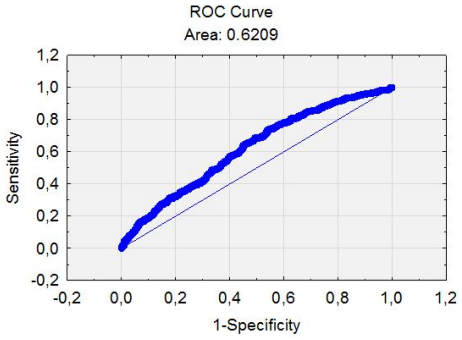


FIGURE F.1: *ROC curve for the grouped dataset.*

TABLE F.3: *Classification ability of the grouped dataset's model calculated from the learning set.*

	Classification of cases Odds ratio: 2,733347 Log odds ratio: 1,005527 Predicted: 1	Predicted: 0	Percent correct
Observed: 1	12	449	2.603
Observed: 0	11	1125	99.032

TABLE F.4: *Odds ratio for the variables modelled with regard to readmission in the grouped dataset.*

Effect	readmitted? - Odds Ratios Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1 Level of Effect	Column	**Odds Ratio**	Lower CL 95%	Upper CL 95%	p
Intercept		1				
Age1		2	**0.9996**	0.99	1.01	0.9537
LOS1		3	**1.0000**	1.00	1.00	0.9736
Area1	5	4	**0.5984**	0.28	1.29	0.5660
Area1	1	5	**0.5846**	0.40	0.85	0.1838
Area1	2	6	**0.6734**	0.50	0.92	0.6473
Area1	3	7	**0.7751**	0.57	1.05	0.4746
Follow up 1	PCH	8	**1.0896**	0.61	1.94	0.3711
Follow up 1	STL_NB_ACT	9	**0.8638**	0.46	1.63	0.3727
Instit1	none	10	**0.4576**	0.22	0.95	0.1762
Instit1	NB	11	**0.4068**	0.19	0.85	0.0424
ICD1	Schizophrenia	12	**2.2407**	1.61	3.13	0.0303
ICD1	SA_Bi_MDDAnx	13	**3.1950**	2.17	4.71	0.0000
ICD1	GMC_Other	14	**1.0186**	0.37	2.80	0.1996
Run 2:						
Area1	4	7	1.7033	1.1704	2.4787	0.0033
Follow up 1	OtherNoneTyg	8	**1.1577**	0.61	2.19	0.9191
Instit1	ACT	11	**2.4580**	1.17	5.16	0.0191
ICD1	SIPD	12	**0.4463**	0.32	0.62	0.0041

F.1.2 Discriminant analysis

The discriminant analysis for the grouped dataset is presented in this section and consists of the significant variables (Table F.5) as well as the classification matrix and ROC curve for both equal prior probabilities (Figure F.2 and Table F.6) and estimated prior probabilities (Figure F.3 and Table F.7).

TABLE F.5: *Summary of all effects of the grouped dataset obtained.*

Effect	Multivariate Tests of Significance [Grouped] Sigma-restricted parameterization Effective hypothesis decomposition					
	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0.984386	25.108259	1	1583	**0.000001**
Follow up 1	Wilks	0.998794	0.955611	2	1583	**0.384799**
"Instit1"	Wilks	0.996070	3.123203	2	1583	**0.044287**
"ICD1"	Wilks	0.975867	13.049058	3	1583	**0.000000**
"Area1"	Wilks	0.993443	2.612012	4	1583	**0.033913**
"Age1"	Wilks	0.999999	0.001559	1	1583	**0.968505**
"LOS1"	Wilks	0.999999	0.001941	1	1583	**0.964869**

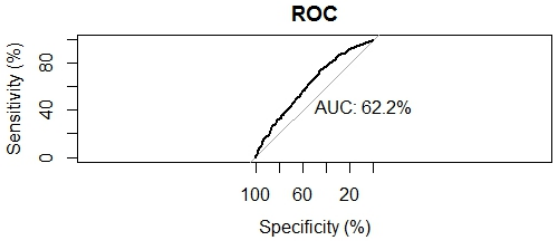


FIGURE F.2: *ROC curve built from the discriminant model in the grouped dataset (equal prior probabilities).*

TABLE F.6: *Classification ability of the discriminant model in the grouped dataset (equal prior probabilities).*

Class	(Classification Matrix [Grouped]) Rows(Observed) Columns(Predicted) PriorProb:Equal		
	Percent Correct	1 p=0.500	0 p=0.500
1	58.13449	268.00000	193.000
0	57.83451	479.00000	657.000
Total	57.92110	747.00000	850.000

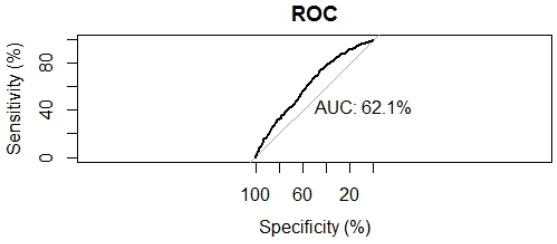


FIGURE F.3: *ROC curve for the discriminant model of the grouped dataset (estimated prior probabilities).*

TABLE F.7: *Classification ability of the discriminant model for the grouped dataset (estimated prior probabilities).*

Class	(Classification Matrix [Grouped]) Rows(Observed) Columns(Predicted) PriorProb:Estimated		
	Percent Correct	1 p=0.2887	0 p=0.7113
1	2.60304	12.00000	449.000
0	99.03169	11.00000	1125.000
Total	71.19599	23.00000	1574.000

F.2 Ungrouped dataset

Similar to the grouped dataset the ungrouped dataset was analysed with both logistic regression and discriminant analysis which is presented in Section F.2.1 and F.2.2 Section respectively.

F.2.1 Logistic regression

The output of the logistic regression analysis for the ungrouped dataset includes: the test of all effects (Table F.4), parameter estimates (Table F.5), odds ratios (Table F.8), ROC curve (Figure F.6) and the classification matrix (Table F.7).

	readmitted? - Test of all effects [LOGIT] Modeled probability readmitted? = 1		
Effect	Degr. of freedom	Wald Stat	p
Intercept	1	14.55197	0.000136
Age1	1	0.01513	0.902103
LOS1	1	0.00454	0.946292
Area1	4	10.27567	0.036032
Follow up 1	6	17.57566	0.007385
ICD1	6	40.65594	0.000000

FIGURE F.4: *'Test of all effects' for the ungrouped dataset.*

	readmitted? - Parameter estimates (VirStatistica in AdmiUngrouped Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1							
Effect	Level of Effect	Column	**Estimate**	Standard Error	Wald Stat	Lower CL 95%	Upper CL 95%	p
Intercept		1	-1.2940	0.3392	14.5520	-1.9588	-0.6291	0.0001
Age1		2	0.0008	0.0063	0.0151	-0.0115	0.0131	0.9021
LOS1		3	0.0001	0.0012	0.0045	-0.0022	0.0024	0.9463
Area1	1	4	-0.2588	0.1504	2.9611	-0.5537	0.0360	0.0853
Area1	2	5	-0.0469	0.1217	0.1487	-0.2854	0.1915	0.6997
Area1	3	6	0.0555	0.1201	0.2139	-0.1799	0.2909	0.6438
Area1	4	7	0.31829	0.126907	6.29040	0.06956	0.567024	0.012139
Follow up 1	PCH	8	0.1315	0.2005	0.4298	-0.2616	0.5245	0.5121
Follow up 1	STL	9	-0.0840	0.2309	0.1322	-0.5365	0.3686	0.7161
Follow up 1	None	10	-0.2342	0.3970	0.3479	-1.0122	0.5439	0.5553
Follow up 1	NB	11	-0.2338	0.2396	0.9519	-0.7035	0.2359	0.3292
Follow up 1	Act	12	0.6498	0.3463	3.5205	-0.0290	1.3286	0.0606
Follow up 1	Other	13	-1.7385	0.9091	3.6572	-3.5203	0.0433	0.0558
ICD1	S	14	0.3328	0.1977	2.8329	-0.0547	0.7203	0.0924
ICD1	SA	15	0.8886	0.2400	13.7122	0.4183	1.3589	0.0002
ICD1	B	16	0.6120	0.2430	6.3424	0.1357	1.0883	0.0118
ICD1	GMC	17	-1.2584	0.9212	1.8660	-3.0640	0.5472	0.1719
ICD1	MDD&Anx	18	0.0694	0.3969	0.0306	-0.7086	0.8473	0.8612
ICD1	Other	19	-0.1878	0.5128	0.1341	-1.1929	0.8173	0.7142
Scale			1.0000	0.0000		1.0000	1.0000	
Run 2:								
Area1	5	6	-0.06806	0.305026	0.04979	-0.66591	0.529776	0.823424
Follow up 1	TYG	13	1.5091	0.5317	8.0556	0.4670	2.5513	0.0045
ICD1	SIPD	17	-0.4565	0.2296	3.9517	-0.9066	-0.0064	0.0468

FIGURE F.5: *Parameter estimates for the regression equation (ungrouped dataset).*

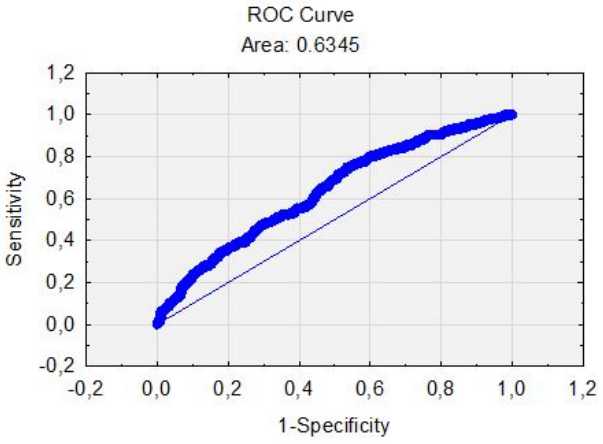


FIGURE F.6: *ROC curve for logistic regression model (ungrouped dataset).*

	Classification of cases Odds ratio: 2,733347 Log odds ratio: 1,005527		
	Predicted: 1	Predicted: 0	Percent correct
Observed: 1	32	429	6.9414
Observed: 0	20	1116	98.239

FIGURE F.7: *Classification ability of the ungrouped dataset's model calculated from the learning set.*

Effect	readmitted? - Odds Ratios Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1 Level of Effect	Column	Odds Ratio	Lower CL 95%	Upper CL 95%	p
Intercept		1.0000				
Age1		2.0000	1.0008	0.9885	1.0132	0.9021
LOS1		3.0000	1.0001	0.9978	1.0024	0.9463
Area1	1.0000	4.0000	0.8263	0.3715	1.8380	**0.0853**
Area1	2.0000	5.0000	1.0214	0.4756	2.1933	0.6997
Area1	3.0000	6.0000	1.1316	0.5263	2.4328	0.6438
Area1	4.0000	7.0000	1.4720	0.6798	3.1855	0.0121
Follow up 1	PCH	8.0000	0.2522	0.0794	0.8009	0.5121
Follow up 1	STL	9.0000	0.2033	0.0619	0.6676	0.7161
Follow up 1	None	10.0000	0.1749	0.0429	0.7129	0.5553
Follow up 1	NB	11.0000	0.1750	0.0530	0.5777	0.3292
Follow up 1	Act	12.0000	0.4234	0.1118	1.6037	**0.0606**
Follow up 1	Other	13.0000	0.0389	0.0037	0.4131	**0.0558**
ICD1	S	14.0000	2.2018	1.5762	3.0756	0.0924
ICD1	SA	15.0000	3.8384	2.4158	6.0988	0.0002
ICD1	B	16.0000	2.9110	1.8141	4.6711	0.0118
ICD1	GMC	17.0000	0.4485	0.0541	3.7175	0.1719
ICD1	MDD&Anx	18.0000	1.6920	0.7094	4.0355	0.8612
ICD1	Other	19.0000	1.3083	0.4154	4.1203	0.7142
Run 2:						
Area1	5	6	1.2102	0.5441	2.6920	0.8234
Follow up 1	TYG	13.0000	3.9657	1.2486	12.5958	0.0045
ICD1	SIPD	17.0000	0.7643	0.2427	2.4072	0.0468

FIGURE F.8: *Odds ratio for the variables in the ungrouped dataset.*

F.2.2 Discriminant analysis

The output of the discriminant analysis for the ungrouped dataset is presented in this section and consists of the significant variables (Table F.8) as well as the classification matrix and ROC curve for both equal prior probabilities (Figure F.9 and Table F.9) and estimated prior probabilities (Figure F.10 and Table F.10).

TABLE F.8: *Summary of all effects - significance of variables in the ungrouped dataset.*

Effect	Multivariate Tests of Significance [Ungrouped] Sigma-restricted parameterization Effective hypothesis decomposition					
	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	1.000000		0		
"Area1"	Wilks	0.993468	2.592121	4	1577	0.035058
Follow up 1	Wilks	0.991008	2.861932	5	1577	0.014047
"Instit1"	Wilks	0.999582	0.659823	1	1577	0.416745
"ICD1"	Wilks	0.973302	7.209533	6	1577	0.000000
"Age1"	Wilks	0.999984	0.025075	1	1577	0.874199
"LOS1"	Wilks	0.999996	0.005787	1	1577	0.939371

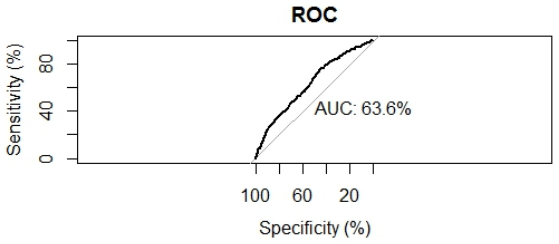


FIGURE F.9: *ROC curve built from the discriminant model of the ungrouped dataset (equal prior probabilities).*

TABLE F.9: *Classification ability of the discriminant model built from the ungrouped dataset (equal prior probabilities).*

Class	(Classification Matrix [Ungrouped]) Rows(Observed) Columns(Predicted) PriorProb:Equal		
	Percent Correct	1 p=0.500	0 p=0.500
1	67.24512	310.00000	151.000
0	51.49648	551.00000	585.000
Total	56.04258	861.00000	736.000

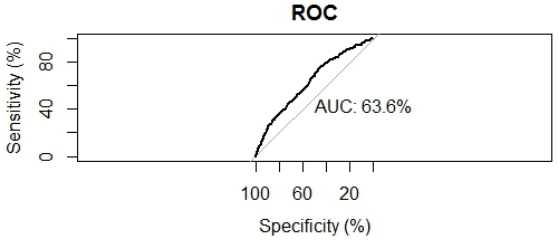


FIGURE F.10: *ROC curve for the discriminant model based on in the ungrouped dataset (estimated prior probabilities).*

TABLE F.10: *Classification ability of the discriminant model built from the ungrouped dataset (estimated prior probabilities).*

Class	(Classification Matrix [Ungrouped]) Rows(Observed) Columns(Predicted) PriorProb:Estimated		
	Percent Correct	1 p=0.2887	0 p=0.7113
1	6.94143	32.00000	429.000
0	98.23944	20.00000	1116.000
Total	71.88478	52.00000	1545.000

F.3 Substance dataset

The substance dataset was briefly analysed with logistic regression and discriminant analysis to predominantly determine if substance use has a significant influence on readmission.

F.3.1 Logistic regression

The logistic regression results pertaining to the substance dataset is presented in Figure F.11 and Tables F.11-F.14. The results are discussed in Section 5.2.1.3 and Section 5.2.1.4. None of the variables emerged as significant and the odds ratios that were significant all contained '1' in their CI[1]:

1. Belonging to neither ACT/NB ('none' group for ACT/NB variable) leads to a patient being 3.4 times more likely to be readmitted than a patient following up at a community program**;

2. Not using tik halves the chance for readmission compared to a patient using tik*;

3. Patients in the STL_NB_ACT follow-up group are three times more likely to be readmitted than patients not from that group**;

4. Patients in the *Tyg_other_none* follow-up group are 3 times less likely to be readmitted than patients not from that group*; and

5. Patients in the *SA_Bi_MDD&Anxiety* group diagnosis group are 2.4 times more likely to be readmitted than patients not from that group*.

TABLE F.11: *'Test of all effects' for the substance dataset.*

Effect	readmitted? - Test of all effects [LOGIT] Modeled probability readmitted? = 1		
	Degr. of freedom	Wald Stat	p
Intercept	1	10.9570	0.0009
Age	1	0.1660	0.6836
LOS	1	0.1650	0.6846
Substance: C	1	0.9034	0.3419
Substance: T	1	2.9939	0.0836
Substance: A	1	0.0039	0.9503
Substance: Other	1	0.0029	0.9568
Substance: none	1	0.6297	0.4275
Area	4	4.4561	0.3478
FollowUp Grouped	2	3.3527	0.1871
ACT/NB	2	5.6253	0.0600
ICD grouped	3	6.0392	0.1097

[1] $*p < 0.1$; $**p < 0.05$; and $***p < 0.01$

TABLE F.12: *Parameter estimates for the regression equation of the substance dataset.*

Effect	readmitted? - Parameter estimates (VirStatistica in Substance_Admi) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1 Level of Effect	Column	**Estimate**	Standard Error	Wald Stat	Lower CL 95%	Upper CL 95%	**p**
Intercept		1	-2.77040	0.836946	10.95696	-4.41078	-1.13001	0.000933
Age		2	0.00594	0.014572	0.16605	-0.02262	0.03450	0.683647
LOS		3	0.00055	0.001348	0.16498	-0.00209	0.00319	0.684609
Substance: C	0	4	-0.23527	0.247526	0.90342	-0.72041	0.24987	0.341867
Substance: T	0	5	-0.33352	0.192755	2.99395	-0.71132	0.04427	0.083576
Substance: A	0	6	-0.01755	0.281541	0.00389	-0.56936	0.53426	0.950293
Substance: Other	0	7	0.01331	0.245574	0.00294	-0.46801	0.49462	0.956792
Substance: none	0	8	-0.25031	0.315422	0.62974	-0.86852	0.36791	0.427450
Area	1	9	-0.63817	0.504215	1.60193	-1.62641	0.35007	0.205631
Area	2	10	0.34156	0.320041	1.13900	-0.28571	0.96883	0.285863
Area	3	11	0.40587	0.331831	1.49602	-0.24451	1.05624	0.221285
Area	4	12	0.55573	0.353949	2.46518	-0.13800	1.24946	0.116395
FollowUp Grouped	PCH	13	0.20970	0.253929	0.68199	-0.28799	0.70739	0.408903
FollowUp Grouped	STL_NB_ACT	14	0.48541	0.280335	2.99822	-0.06404	1.03485	0.083356
ACT/NB	none	15	0.80358	0.372567	4.65207	0.07336	1.53380	0.031016
ACT/NB	NB	16	-0.38576	0.419036	0.84749	-1.20706	0.43553	0.357264
ICD grouped	Schizoprehia	17	0.39414	0.349408	1.27245	-0.29069	1.07897	0.259308
ICD grouped	SIPD	18	-0.78323	0.494000	2.51379	-1.75145	0.18499	0.112854
ICD grouped	SA_Bi_MDDAnx	19	0.63861	0.369184	2.99217	-0.08498	1.36220	0.083668
Scale			1.00000	0.000000		1.00000	1.00000	
Run 2:								
Area1	5	12	-0.66499	0.882322	0.56804	-2.39431	1.06433	0.451040
FollowUp Grouped	none_other_tyg	13	-0.69511	0.403080	2.97389	-1.48513	0.09491	0.084618
ACT/NB	ACT	15	-0.41782	0.569099	0.53901	-1.53323	0.69760	0.462844
ICD grouped	GMC_Oth	17	-0.24952	0.855252	0.08512	-1.92578	1.42675	0.770479

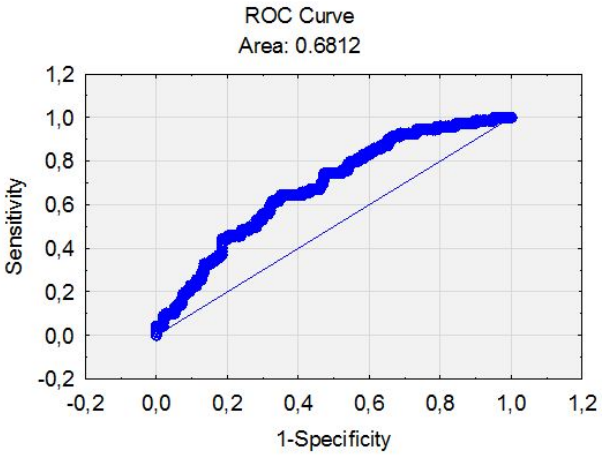


FIGURE F.11: *ROC curve for logistic regression model based on the substance dataset.*

TABLE F.13: *Classification ability of the substance dataset's model calculated from the learning set.*

	Classification of cases		
	Predicted: 1	Predicted: 0	Percent correct
Observed: 1	2	68	2.8571
Observed: 0	0	228	100

TABLE F.14: *Odds ratio for the variables in the substance dataset.*

Effect	readmitted? - Odds Ratios Distribution : BINOMIAL, Link function: LOGIT Modeled probability that readmitted? = 1 Level of Effect	Column	**Odds Ratio**	Lower CL 95%	Upper CL 95%	**p**
Intercept		1				
Age		2	1.0060	0.9776	1.0351	0.6836
LOS		3	1.0005	0.9979	1.0032	0.6846
Substance: C	0	4	0.6247	0.2367	1.6483	0.3419
Substance: T	0	5	0.5132	0.2411	1.0926	0.0836
Substance: A	0	6	0.9655	0.3202	2.9111	0.9503
Substance: Other	0	7	1.0270	0.3922	2.6892	0.9568
Substance: none	0	8	0.6062	0.1760	2.0872	0.4275
Area	1	9	1.0272	0.0939	11.2331	0.2056
Area	2	10	2.7361	0.3113	24.0464	0.2859
Area	3	11	2.9179	0.3243	26.2521	0.2213
Area	4	12	3.3896	0.3667	31.3303	0.1164
FollowUp Grouped	PCH	13	2.4715	0.7438	8.2116	0.4089
FollowUp Grouped	STL_NB_ACT	14	3.2561	0.9175	11.5554	0.0834
ACT/NB	none	15	3.3919	0.6214	18.5137	0.0310
ACT/NB	NB	16	1.0326	0.1677	6.3584	0.3573
ICD grouped	Schizoprehia	17	1.9034	0.2036	17.7954	0.2593
ICD grouped	SIPD	18	0.5864	0.0516	6.6684	0.1129
ICD grouped	SA_Bi_MDDAnx	19	2.4306	0.2556	23.1126	0.0837
Run 2:						
Area1	5	12	0.9735	0.0890	10.6464	0.4510
FollowUp Grouped	none_other_tyg	13	0.3071	0.0865	1.0899	0.0846
ACT/NB	ACT	15	0.9685	0.1573	5.9636	0.4628
ICD grouped	GMC_Oth	17	0.4114	0.0433	3.9123	0.7705

F.3.2 Discriminant analysis

The results of the discriminant analysis pertaining to the substance dataset is presented in Figures F.12-F.13 and Tables F.15-F.17.

TABLE F.15: *Summary of all effects - significance of variables in the substance dataset.*

Effect	Multivariate Tests of Significance [Substance] Sigma-restricted parameterization Effective hypothesis decomposition Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0.999981	0.005172	1	279	0.942719
Substance: C	Wilks	0.996699	0.923962	1	279	0.337269
Substance: T	Wilks	0.989509	2.958065	1	279	0.086559
Substance: A	Wilks	0.999975	0.007030	1	279	0.933241
Substance: Other	Wilks	0.999993	0.001847	1	279	0.965754
Substance: none	Wilks	0.998089	0.534226	1	279	0.465449
Area	Wilks	0.983299	1.184668	4	279	0.317735
FollowUp Grouped	Wilks	0.986299	1.937823	2	279	0.145951
ACT/NB	Wilks	0.978940	3.001092	2	279	**0.051341**
ICD grouped	Wilks	0.977095	2.180127	3	279	**0.090596**
Age	Wilks	0.999413	0.163984	1	279	0.685825
LOS	Wilks	0.999416	0.163165	1	279	0.686569

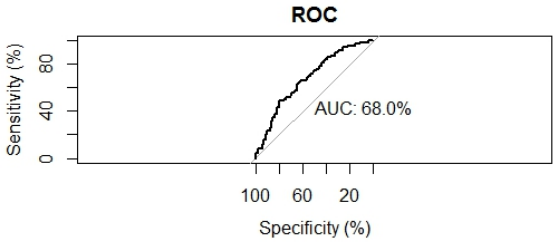


FIGURE F.12: *ROC curve for the discriminant model built from the substance dataset (equal prior probabilities).*

TABLE F.16: *Classification ability of the discriminant model built from the substance dataset (equal prior probabilities).*

Class	(Classification Matrix [Substance]) Rows(Observed) Columns(Predicted) PriorProb:Equal Percent Correct	1 p=0.500	0 p=0.500
1	68.57143	48.00000	22.000
0	55.70175	101.00000	127.000
Total	58.72483	149.00000	149.000

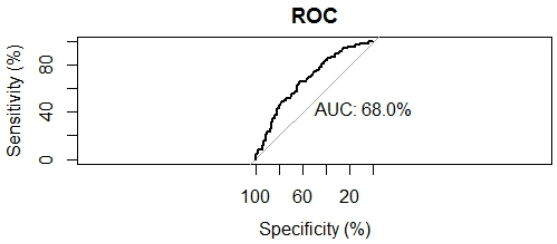


FIGURE F.13: *ROC curve for the discriminant model based on estimated prior probabilities in the substance dataset.*

TABLE F.17: *Classification ability of the discriminant model based on estimated prior probabilities in the substance dataset.*

Class	(Classification Matrix [Substance] Rows(Observed) Columns(Predicted) PriorProb:Estimated Percent Correct	1 p=0.2349	0 p=0.7651
1	4.28571	3.00000	67.000
0	99.12281	2.00000	226.000
Total	76.84564	5.00000	293.000

APPENDIX G

CART analysis output

This appendix consists of tables and figure depicting the output of the CART analyses on both the grouped and ungrouped dataset.

G.1 Grouped dataset

Tree 50 and Tree 51 which respectively have three and two terminal nodes were chosen as the best models for predicting readmissions. The importance graph (Figures G.1 and G.3), classification table (Tables G.1 and G.2) and categorised histogram (Figures G.2 and G.4) for both models are presented in this section.

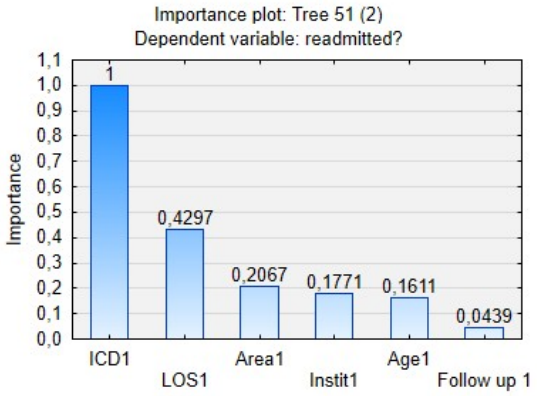


FIGURE G.1: *Variable importance as calculated by CART at Tree 51*

TABLE G.1: *Classification ability of the CART model for Tree 51*

	Classification matrix 51 (Grouped Dataset) Dependent variable: readmitted? Options: Categorical response			
	Observed	Predicted 0	Predicted 1	Row Total
Number	0	294	842	1136
Column Percentage		83.52%	67.63%	
Row Percentage		25.88%	74.12%	
Total Percentage		18.41%	52.72%	71.13%
Number	1	58	403	461
Column Percentage		16.48%	32.37%	
Row Percentage		12.58%	87.42%	
Total Percentage		3.63%	25.23%	28.87%
Count	All Groups	352	1245	1597
Total Percent		22.04%	77.96%	

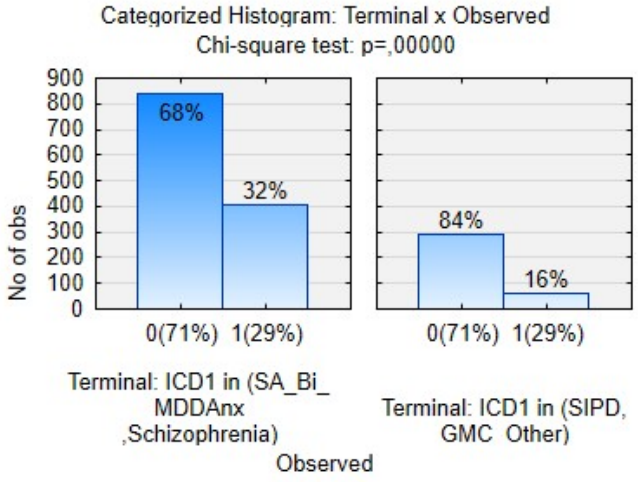


FIGURE G.2: *Prediction for patients readmitted (1) or not (0) in the terminal nodes of Tree 51*

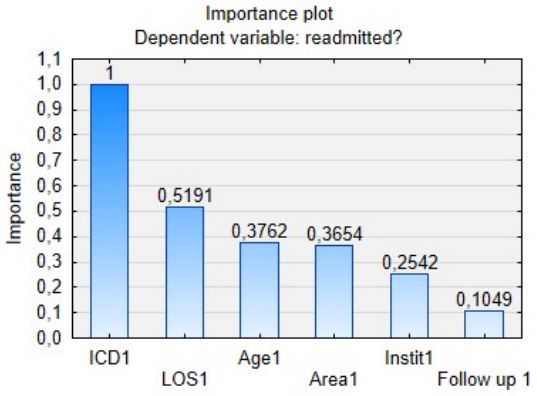


FIGURE G.3: *Variable importance as calculated by CART at Tree 50*

TABLE G.2: *Classification ability of the CART model for Tree 50*

	Classification matrix 50 (Grouped Dataset) Dependent variable: readmitted? Options: Categorical response			
	Observed	Predicted 0	Predicted 1	Row Total
Number	0	342	794	1136
Column Percentage		83.62%	66.84%	
Row Percentage		30.11%	69.89%	
Total Percentage		21.42%	49.72%	71.13%
Number	1	67	394	461
Column Percentage		16.38%	33.16%	
Row Percentage		14.53%	85.47%	
Total Percentage		4.20%	24.67%	28.87%
Count	All Groups	409	1188	1597
Total Percent		25.61%	74.39%	

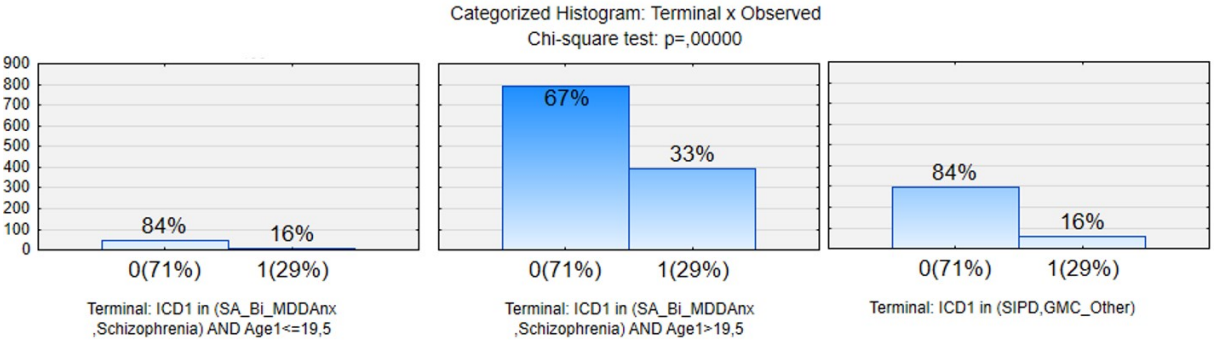


FIGURE G.4: *Prediction for patients readmitted (1) or not (0) in the terminal nodes of Tree 50*

G.2 Ungrouped dataset

The ungrouped dataset was also evaluated with regard to the CART models. The three trees with the fewest terminal nodes and CV costs are displayed and partially discussed in this section.

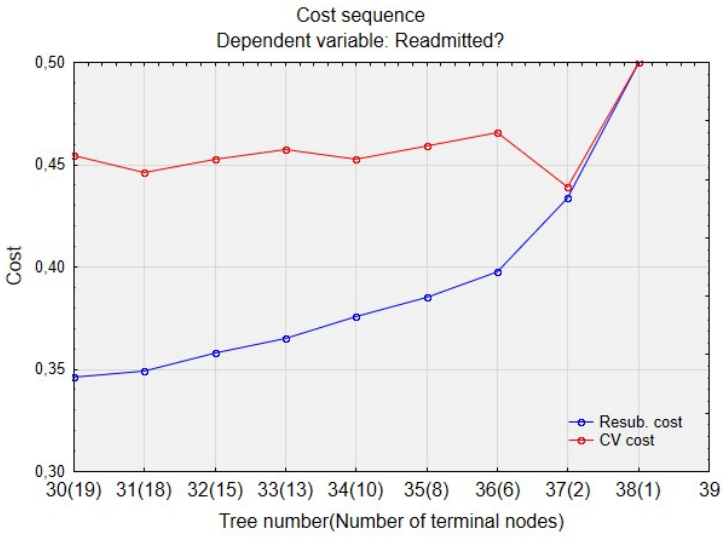


FIGURE G.5: *Cost sequence diagram for the CART model (zoomed in)*

G.2.1 Tree 37 (2 terminal nodes)

This section pertains to Tree 37 which was chosen the best tree owing to having a small CV cost and only two terminal nodes compared to Tree 36 which has six terminal nodes and a higher CV cost. The importance plot for Tree 37 and the classification ability of the model is respectively displayed in Figure G.6 and Figure G.3. The decision rule categorised histogram is displayed in Figure G.7.

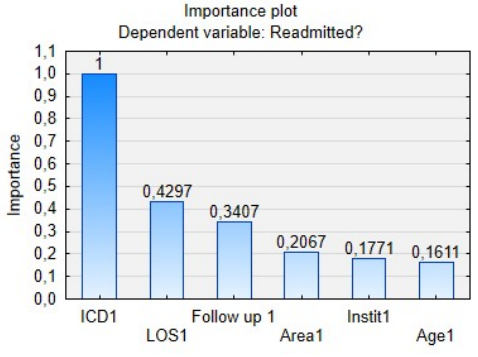


FIGURE G.6: *Variable importance as calculated by CART at Tree 37*

TABLE G.3: *Classification ability of the CART model for Tree 37*

	Classification matrix 37 Dependent variable: Readmitted? Options: Categorical response			
	Observed	Predicted 1	Predicted 0	Row Total
Number	1	403	58	461
Column Percentage		32.37%	16.48%	
Row Percentage		87.42%	12.58%	
Total Percentage		25.23%	3.63%	28.87%
Number	0	842	294	1136
Column Percentage		67.63%	83.52%	
Row Percentage		74.12%	25.88%	
Total Percentage		52.72%	18.41%	71.13%
Count	All Groups	1245	352	1597
Total Percent		77.96%	22.04%	

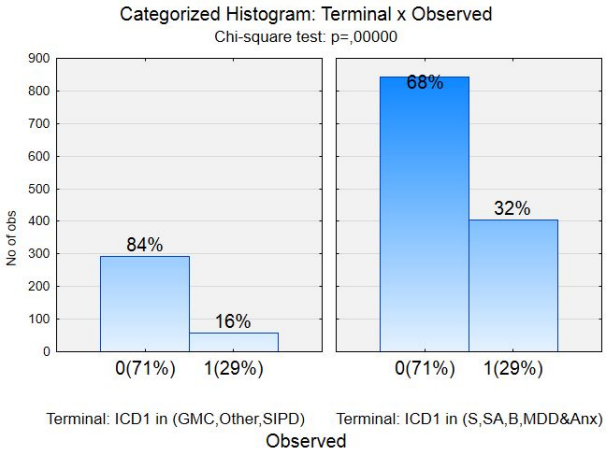


FIGURE G.7: *Prediction for patients being readmitted (1) or not (0) in the terminal node classes of Tree 37*

G.2.2 Tree 36 (6 terminal nodes)

Tree 36 has the least terminal nodes after Tree 37 (excluding tree 38 owing to a high CV cost), but a much higher CV cost and Tree 35, although having a smaller CV cost than Tree 37 has eight terminal nodes, which suggest a over fitted tree. For example, the third picture from left in the top row specifies that a patient with either *schizophrenia*, *schizo-affective*, *bipolar*,*MDD or anxiety* **and** who follows up at either a *clinic*, *Stikland*, *nowhere*, *New Beginnings* or *other* **and** is younger than 19.5 years have a 14% chance to be readmitted. This rule is already very specific and some rules might be based on randomness or noise which suggest that the model will not perform well with classifying new data.

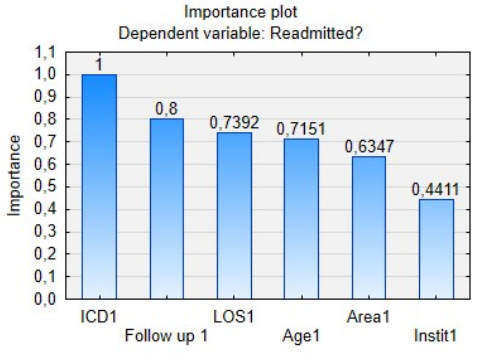


FIGURE G.8: *Variable importance as calculated by CART at Tree 36*

TABLE G.4: *Classification ability of the CART model for Tree 36*

	Classification matrix 36 Dependent variable: Readmitted? Options: Categorical response			
	Observed	Predicted 1	Predicted 0	Row Total
Number	1	276	185	461
Column Percentage		38.17%	21.17%	
Row Percentage		59.87%	40.13%	
Total Percentage		17.28%	11.58%	28.87%
Number	0	447	689	1136
Column Percentage		61.83%	78.83%	
Row Percentage		39.35%	60.65%	
Total Percentage		27.99%	43.14%	71.13%
Count	All Groups	723	874	1597
Total Percent		45.27%	54.73%	

The predictive capability of the three tree models along with the CV cost and amount of terminal nodes are summarised in Table G.5. Tree 35 on average classifies 61.5% of the cases correctly which is the best of the three trees. The tree however have multiple splits and as previously discussed might not fare well with predicting new data owing to being over-fitted. Accordingly, Tree 38 is chosen as the best tree.

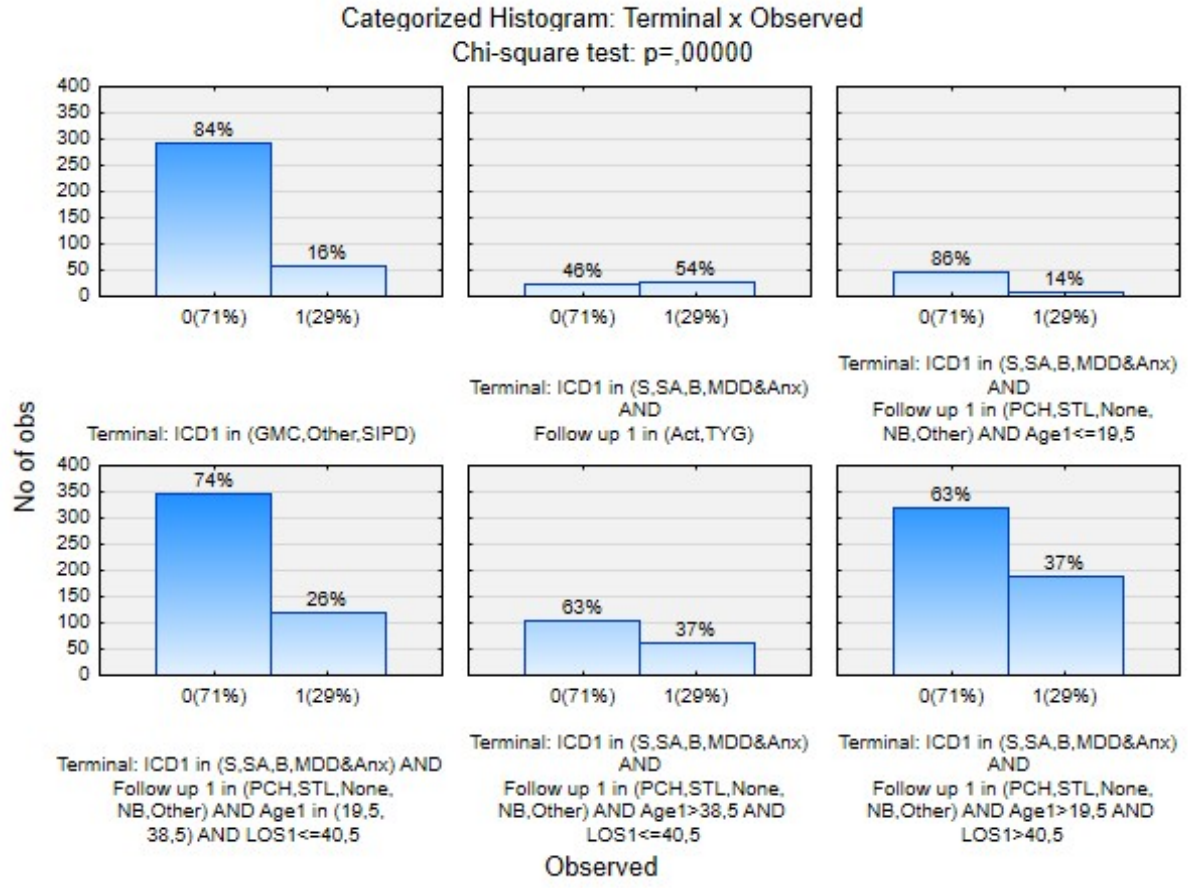


FIGURE G.9: *Categorised histogram for the terminal nodes and chance for readmission = 1 or 0 according to the splitting rules for the node*

TABLE G.5: *Summary of the prediction capability of Tree 35, 36 and 37 respectively*

	Tree 37	Tree 36	Tree 35
% classified correct			
1	87.4%	59.9%	64.2%
0	25.9%	60.7%	58.7%
average	56.7%	60.3%	61.5%
Other information			
# terminal nodes	2	6	8
CV cost	0.43	0.47	0.46